The statistical work of Lucien Le Cam

van der Vaart, A.W.

published in
Annals of Statistics
2002

DOI (link to publisher)
10.1214/aos/1028674836

document version
Publisher's PDF, also known as Version of record

The Annals of Statistics
2002, Vol. 30, No. 3, 631–682

THE STATISTICAL WORK OF LUCIEN LE CAM

BY AAD VAN DER VAART

Free University, Amsterdam

We give an overview and appraisal of the scientific work in theoretical statistics, and its impact, by Lucien Le Cam. The references to Le Cam's papers refer to the Le Cam bibliography. The reference is the first paper for the given year if not stated.

1. Introduction. Lucien Le Cam died in April 2000 leaving more than 80 articles and several books. This paper is an attempt to give a review of his contributions to statistics and their impact on present-day statistics. As is well known these contributions are many, and they are not always easy to consume in their original form. Le Cam's 1986 book can be viewed as his own summary of his theory, up to that date, and has a reputation of being hard to read. Many of Le Cam's papers are equally deep and may become fully appreciated only in the future. Of necessity therefore this review is an eclectic one: we can only review what we think we understand. Furthermore, our review will be biased towards what we think is most important from the present-day point of view.

We make this point not just for modesty and politeness, but because there are good reasons to believe that the appreciation of Le Cam's work may still change. For instance, Le Cam's theory of comparison of experiments, based on a distance between statistical experiments, while being one of his central contributions, has attracted more interest in the last ten years than before. This theory was at the core of Le Cam's thinking, and many of his discoveries were made from the framework of comparison of experiments. In contrast, typical applications, for instance local asymptotic normality, have become popular in a more direct fashion. The future may see a different perspective.

We have deliberately put the adjective "statistical" in the title of this paper. Le Cam's main scientific work is on mathematical statistics, but he has also contributed significantly to probability theory and other subjects. This includes work on the central limit theorem, empirical processes, Poissonization, weak convergence theory and the history of mathematics. While it is not clear from the title, we shall also not discuss Le Cam's applied statistical work.

The book Le Cam (1986) takes its point of departure in the very abstract, by describing a statistical experiment as a subset of a Riesz lattice. This abstraction was not a late career synthesis, but was already fully present in one of his earliest

Received July 2001; revised October 2001.

AMS 2000 subject classifications. 62G15, 62G20, 62F25.

Key words and phrases. Limit experiment, deficiency, LAN, contiguity, metric entropy, comparison of experiments, sufficiency.

papers, Le Cam (1964a). Nevertheless, the strategy of this review is different: we start with topics that have become fairly familiar, and end with the more abstract structures.

In doing so we fall into several traps that Le Cam warned against. Initially we shall be discussing limit results, rather than approximation results. Furthermore, we shall use the case of i.i.d. observations as the main example.

Asymptotic statistics is often equated to "limit theorems." One of Le Cam's achievements was to connect these to approximation results. Regarding limit theorems Le Cam [(1986), page xiv] writes

> From time to time results are stated as limit theorems obtainable as something called n "tends to infinity." This is especially so in Chapter 7 where the results are just limit theorems. Otherwise we have made a special effort to state the results in such a way that they could eventually be transformed into approximation results. Indeed, limit theorems "as n tends to infinity" are logically devoid of content about what happens at any particular n. All they can do is suggest certain approaches whose performance must then be checked on the case at hand. Unfortunately the approximation bounds we could get were too often too crude and cumbersome to be of any practical use. Thus we have let n tend to infinity, but we would urge the reader to think of the material in approximation terms, especially in subjects such the ones described in Chapter 11.

Chapter 11 is about the construction of asymptotically efficient estimators in LAN models, roughly variations on maximum likelihood estimators that are shown to be asymptotically normal.

On the example of i.i.d. observations Le Cam [(1986), pages 555–556] comments

> However, since it is the standard i.i.d. case with its quaint concepts, such as "consistency," that occupies so much of the literature, we have devoted this chapter to it in order to illustrate the applicability of the general ideas of the present volume.

In this review we shall also find the example useful to illustrate ideas that are otherwise obscured by technical details.

The point of departure for Le Cam's thinking about statistics was Wald's "theory of statistical functions" [see Wald (1950)]. This assumes given a set of probability measures $(P_\theta : \theta \in \Theta)$ on some measurable space, a decision space and a loss function. The set of probability measures can be called a "statistical model." Le Cam usually preferred the word "experiment," a term that he derived from Blackwell (1951), even though it was already familiar to Wald (1939) [Le Cam (1986), page xvi]. The word "experiment" appears to have become closely connected to Le Cam's work: carrying out research on statistical experiments is often understood as being equivalent to "working on Le Cam type theory."

As we shall see, Le Cam coined the word "experiment" for more general sets than sets of probability measures. In this paper an *experiment* $\mathcal{E} = (\mathcal{X}, \mathcal{A}, P_\theta : \theta \in \Theta)$ will generally be understood to be an indexed set of probability distributions $(P_\theta : \theta \in \Theta)$ on a measurable space $(\mathcal{X}, \mathcal{A})$. Le Cam's definition of an experiment is given in Section 8.

2. Local asymptotic normality. "Locally asymptotically normal families of distributions" is the title of a major paper by Le Cam, published in the University of California Publications series of 1960. The concept of *local asymptotic normality* is probably among Le Cam's best-known contributions and is also referred to by the acronym LAN. This acronym does not appear in the paper, the closest relatives being the letter combinations DN and DAN for "Differentially (Asymptotically) Normal."

The 1960 paper [Le Cam (1960a)] starts with a tribute to J. Neyman:

> The present paper is an outcome of conversations between Professor J. Neyman and the author about the construction of asymptotically similar tests. The adjective "asymptotically" is used to convey two ideas. First, the information provided by the sample is sufficient to give very sharp estimates of the parameters involved. Second, in the range of "probable" values of these estimates, the family of probability measures under study can be approximated very closely by a family of a simpler nature.

From a historical perspective the last sentence is interesting, because it shows that local asymptotic normality was conceived by Le Cam from the beginning as a way of approximating statistical experiments. In contrast, the concept became popular as a formalization of a Taylor type expansion of a likelihood function around a fixed, true parameter. A simplified version of the definition by Le Cam and Yang (1990), which is of the latter type, is as follows.

For each n let $(P_{n,\theta} : \theta \in \Theta)$ be an indexed family of probability measures on some measurable space $(\mathcal{X}_n, \mathcal{A}_n)$, where Θ is an open subset of $\mathbb{R}^k$. Let δ_n be positive numbers with $\delta_n \to 0$. This family is called LAN at $\theta \in \Theta$ if there exist a sequence of stochastic vectors $\Delta_{n,\theta}$ and a nonsingular $(k \times k)$ matrix J_θ such that $\Delta_{n,\theta} \rightsquigarrow N(0, J_\theta)$ under $P_{n,\theta}$ and such that for every bounded sequence of vectors h_n,

$$\log \frac{dP_{n,\theta+\delta_n h_n}}{dP_{n,\theta}} - h_n^T \Delta_{n,\theta} + \frac{1}{2} h_n^T J_\theta h_n \overset{P_{n,\theta}}{\to} 0 \tag{2.1}$$

(we use the wiggly arrow $\rightsquigarrow$ to denote convergence in distribution of random elements in metric spaces). The word "normality" in LAN may be explained from the asymptotic normality of the sequence $\Delta_{n,\theta}$, but is better understood in a different way, as we shall see.

During the last 40 years many statistical models were encountered where the LAN concept plays a central role. The best-known example is that of replicated experiments in which the distribution of a single observation depends smoothly on a Euclidean parameter. Specifically, the sequence of experiments in which $P_{n,\theta} = P_\theta^n$ is the distribution of an i.i.d. sample $X_1, \ldots, X_n$ from a density p_θ such that the map $\theta \mapsto p_\theta$ is differentiable [in the precise sense of (12.1) below] is LAN, with $\delta_n = 1/\sqrt{n}$ and the "centering sequence" $\Delta_{n,\theta}$ and "Fisher information matrix" J_θ determined from the score function $\dot{\ell}_\theta$ of the model through

$$\Delta_{n,\theta} = \frac{1}{\sqrt{n}} \sum_{i=1}^{n} \dot{\ell}_\theta(X_i), \qquad J_\theta = P_\theta \dot{\ell}_\theta \dot{\ell}_\theta^T.$$

[See (12.1) for a definition of the score function.] Not uncharacteristically for Le Cam's writings, which often ask attention for unusual cases, this example is contained in his 1960 paper, but as a third example, not in the prominent place it deserved. In the 1960 paper the elegant way of defining "differentiability" of a density in the parameter by employing the root $\theta \mapsto p_\theta^{1/2}$ (see Section 12 below) does not appear yet, but conditions are stated in terms of higher order derivatives. For sufficiently regular parametric families $\theta \mapsto p_\theta$ it is not difficult to derive the LAN expansion from an ordinary Taylor expansion, much in the way that a standard analysis of maximum likelihood estimators proceeds.

A Taylor expansion of the log likelihood is not the main point of Le Cam's 1960 paper. He uses the expansion to show the existence of asymptotically normal estimators that are "asymptotically sufficient." Using these estimators he is able to show that certain decision procedures, based on normal approximations, are asymptotically optimal.

The idea is as follows. We can write the LAN assumption (2.1) in the form

$$dP_{n,\theta+\delta_n h} = \exp\{h^T \Delta_{n,\theta} - \tfrac{1}{2} h^T J_\theta h + \cdots\}\, dP_{n,\theta}. \tag{2.2}$$

If we consider θ as known, think of h as parametrizing the model, and ignore the remainder term $\cdots$, then we see that the likelihood (relative to the dominating measure $P_{n,\theta}$) depends on the data only through the statistic $\Delta_{n,\theta}$. In other words, the statistic $\Delta_{n,\theta}$ is sufficient in the statistical model $(P_{n,\theta+\delta_n h} : h \in \mathbb{R}^k)$, for fixed θ.

In view of the remainder term, the sufficiency can only be true in an asymptotic sense, in general. It is instructive to write down the preceding display for the case of observing n i.i.d. observations $X_1, \ldots, X_n$ from the $N(\theta, 1)$ distribution. For $\delta_n = n^{-1/2}$ this takes the form

$$\prod_{i=1}^n \phi(X_i - \theta - \delta_n h) = \exp\left\{h\delta_n \sum_{i=1}^n (X_i - \theta) - \tfrac{1}{2}h^2\right\} \prod_{i=1}^n \phi(X_i - \theta).$$

In this case the remainder term vanishes, $J_\theta = 1$, and the centering sequence

$$\Delta_{n,\theta} = \frac{1}{\sqrt{n}} \sum_{i=1}^n (X_i - \theta)$$

is exactly sufficient. This centering variable is exactly $N(h, 1)$-distributed under $\theta + \delta_n h$. By the sufficiency the experiment $(N(h, 1) : h \in \mathbb{R})$ corresponding to observing $\Delta_{n,\theta}$ is statistically equivalent to the original experiment. We shall see, more generally, that for an LAN sequence of experiments the experiment $(P_{n,\theta+\delta_n h} : h \in \mathbb{R}^k)$ is "asymptotically equivalent" to the experiment $(N(h, J_\theta^{-1}) : h \in \mathbb{R}^k)$.

The model $(P_{n,\theta+\delta_n h} : h \in \mathbb{R}^k)$ is a "local experiment" indexed by a "local parameter" h. It depends on the "original parameter" θ, as do the "sufficient

statistics" $\Delta_{n,\theta}$. Therefore, the preceding observations may seem to carry little relevance. However, good estimators T_n of θ ought to satisfy

$$\sqrt{n}(T_n - \theta) - J_\theta^{-1}\Delta_{n,\theta} \overset{P_{n,\theta}}{\to} 0. \tag{2.3}$$

In current terminology, following Hájek (1970), an estimator sequence T_n with such a property is called "best regular." It is also "asymptotically efficient" in the sense of Rao [(1965), page 285]. Even though Le Cam (1960a) refrained from arguing so, the method of maximum likelihood yields a best regular sequence, under some conditions. Many Bayes estimators are best regular as well, under some conditions. Le Cam (1960a) constructed a particular sequence T_n satisfying (2.3) under minimal conditions.

If again we ignore the remainder term, then for every given θ and n, the statistic T_n in (2.3) is an affine function of $\Delta_{n,\theta}$ and hence it is statistically equivalent to $\Delta_{n,\theta}$. This and the "asymptotic sufficiency" of $\Delta_{n,\theta}$ suggest that the sequence of statistics $\sqrt{n}(T_n - \theta)$, or equivalently T_n, is "asymptotically sufficient" in the local experiments $(P_{n,\theta+\delta_n h} : h \in \mathbb{R}^k)$, for every fixed θ. Because T_n does not depend on θ, Le Cam would write that they are *asymptotically sufficient* in the original experiments $(P_{n,\theta} : \theta \in \Theta)$.

Le Cam (1960a) succeeded in making this reasoning precise in the following way. Call a sequence of experiments $(Q_{n,\theta} : \theta \in \Theta)$ *differentially asymptotically equivalent* to the experiments $(P_{n,\theta} : \theta \in \Theta)$ if for every n the sample spaces are the same and for all θ and every compact set K,

$$\sup_{h \in K} \| P_{n,\theta+\delta_n h} - Q_{n,\theta+\delta_n h} \| \to 0.$$

The norm $\|\cdot\|$ is the total variation norm, which can be defined as, with the supremum taken over all measurable functions f,

$$\|Q\| = \sup_{\|f\|_\infty \le 1} |Qf|. \tag{2.4}$$

This is a very strong norm, and therefore the local experiments $(P_{n,\theta+\delta_n h} : h \in K)$ and $(Q_{n,\theta+\delta_n h} : h \in K)$ for a given pair of differentially asymptotically equivalent experiments ought to be equivalent in a statistical sense. For the testing problem, which Le Cam (1960a) set out to solve, the equivalence is immediate from the definition of the total variation norm: for every sequence of test functions ϕ_n the difference between its power functions $h \mapsto P_{n,\theta+\delta_n h}\phi_n$ and $h \mapsto Q_{n,\theta+\delta_n h}\phi_n$ in the two local experiments tends to zero, uniformly in $h \in K$. Thus the two sequences of experiments are equivalent for testing in the sense of allowing approximately the same power functions. The equivalence can be shown to extend to all statistical decision problems with bounded loss functions.

"Asymptotic sufficiency" of statistics will now be understood in a precise mathetical sense as being exactly sufficient in *some* sequence of differentially asymptotically equivalent experiments. The following theorem says that for estimators T_n

satisfying (2.3) such differentially asymptotically equivalent experiments can always be found.

The theorem is in the spirit of Le Cam's (1960a) Theorem 5.1 on page 58. In that theorem Le Cam uses special estimators T_n, whereas the following theorem allows general estimators satisfying (2.3). In this more general form the theorem is also due to Le Cam, but it is only informally stated in Le Cam (1960a), in Appendix B of this paper, written to answer three questions "raised by a reader of this paper in manuscript" [Le Cam (1960a), page 94]. Condition (2.3) is precisely Le Cam's property (P) on page 94.

THEOREM 2.1. *Assume that the LAN condition* (2.1) *holds*, *that the map* $\theta \mapsto P_{n,\theta}(A)$ *is measurable for every measurable set* A, *and that the map* $\theta \mapsto J_\theta^{-1}$ *is continuous*. *Let* T_n *satisfy* (2.3). *Then there exists a sequence of experiments* $(Q_{n,\theta} : \theta \in \Theta)$ *that is differentially asymptotically equivalent to* $(P_{n,\theta} : \theta \in \Theta)$ *such that* T_n *is sufficient in* $(Q_{n,\theta} : \theta \in \Theta)$ *for every fixed* n.

We have already indicated that the maximum likelihood estimator is a typical candidate for T_n. However, this estimator will only satisfy (2.3) under restrictive regularity conditions. Another major result of Le Cam (1960a) is a recipe for the construction of estimators T_n that satisfy (2.3). For this the LAN condition is not sufficient, because this condition concerns the original experiments only in a local sense. To make the connection between LAN and the global problem, Le Cam's (1960a) conditions (DN) also include the requirement that there exist estimators $\hat{\theta}_n$ that are δ_n^{-1}-consistent. These are estimators such that the sequence $\delta_n^{-1}(\hat{\theta}_n - \theta)$ is uniformly tight under $P_{n,\theta}$, for every $\theta \in \Theta$. Then Le Cam (1960a) proposes to estimate θ by

$$T_n = \hat{\theta}_n^* + \delta_n J_{\hat{\theta}_n^*}^{-1} \Delta_{n,\hat{\theta}_n^*}, \qquad \hat{\theta}_n^* = \hat{\theta}_n + \delta_n v_n,$$

for v_n variables that are uniform on the unit cube, for instance, and $\Delta_{n,\theta}$ statistics as in the LAN expansion (2.1), but subject to some minor continuity requirements. [Because (2.1) is an approximation, the statistics $\Delta_{n,\theta}$ are not uniquely determined; Le Cam shows that one can always choose appropriate versions.] These estimators possess the desired limit behavior (2.3) under minimal conditions, and hence are alternatives to the maximum likelihood estimator.

The noise v_n added to the preliminary estimators $\hat{\theta}_n$ has the purpose of avoiding the need of differentiability of the maps $\theta \mapsto \Delta_{n,\theta}$. An alternative would be to discretize the preliminary estimators to a grid of meshwidth δ_n [see Le Cam (1969) or Le Cam and Yang (1990)]. Both methods yield estimators $\hat{\theta}_n^*$ "which do not search for the singularities of $\Delta_{n,\theta}$" [Le Cam (1960a), page 95], unlike maximum likelihood estimators.

The preceding theorem implies immediately that for every sequence of tests ϕ_n in $(P_{n,\theta} : \theta \in \Theta)$ there exists a sequence of tests ψ_n based on T_n such that, for every θ and every compact set $K \subset \mathbb{R}^k$,

$$\sup_{h \in K} |P_{n,\theta+\delta_n h}\phi_n - P_{n,\theta+\delta_n h}\psi_n| \to 0.$$

Indeed, by the differential asymptotic equivalence this is true if it is true for the $Q_{n,\theta+\delta_n h}$ instead of the $P_{n,\theta+\delta_n h}$, and for the $Q_{n,\theta+\delta_n h}$ the left side of the display can be reduced to zero, by the sufficiency of T_n, for every n.

On the other hand, the theorem does not immediately tell us what types of asymptotic power functions are possible in the original experiments. By the "contiguity arguments" explained in the next section it can be seen that (2.3) implies that the sequence $\delta_n^{-1}(T_n - \theta)$ is asymptotically $N(h, J_\theta^{-1})$-distributed under $P_{n,\theta+\delta_n h}$, for every h. This and the sufficiency of each T_n suggest that, asymptotically, the possible power functions in the local experiments $(P_{n,\theta+\delta_n h} : h \in \mathbb{R}^k)$ are the power functions that are possible in the experiment $(N(h, J_\theta^{-1}) : h \in \mathbb{R}^k)$. Le Cam makes this intuition rigorous in the last section of Le Cam (1960a), and uses it to study the power functions of asymptotically similar tests. We shall discuss the relationship between the experiments $(P_{n,\theta+\delta_n h} : h \in \mathbb{R}^k)$ and $(N(h, J_\theta^{-1}) : h \in \mathbb{R}^k)$ in Section 5 within the more general context of (weak) convergence of experiments, introduced by Le Cam (1972a). The sequence of statistics T_n satisfying (2.3) will then, besides asymptotically sufficient, also be seen to be *distinguished*: its set of limit distributions $(N(h, J_\theta^{-1}) : h \in \mathbb{R}^k)$ is a limit experiment for the sequence of experiments $(P_{n,\theta+\delta_n h} : h \in \mathbb{R}^k)$. The distinguishedness allows to formalize and generalize conclusions concerning optimality of certain "asymptotically normal procedures." In the 1960 paper the asymptotic sufficiency appears to be the main point. The potential application to asymptotic optimality is mentioned, but almost in passing [e.g., Le Cam (1960a), page 84, paragraph 4], except for the testing case.

Thus the idea that local asymptotic normality concerns approximation by a normal experiment, not just an expansion of a likelihood ratio process is present already in 1960. It became more explicit later. Le Cam (1964a) introduced the deficiency measure, and Le Cam (1969) established the relationship between the convergence of likelihood ratios (2.1) and convergence in the deficiency distance. (See Sections 5 and 7.) Approximation in the deficiency distance can be viewed as a strengthening of the asymptotic sufficiency argument of Theorem 2.1, which also immediately characterizes the set of available asymptotic risk functions.

A full description of locally asymptotically normal experiments requires a localization through initial estimators $\hat{\theta}_n$, and the validity of a Taylor-type expansion (2.1) of the log likelihood ratios. The power of the concept is that once these two requirements are fulfilled statistical questions can be answered in a unified way. It is not necessary to impose further structure, such as the structure offered by i.i.d. observations.

Local asymptotic normality was subsequently established for other statistical problems. For instance:

- reduction to the observation of sample moments
- observation of stationary Markov chains with a transition density depending smoothly on the parameter
- Gaussian time series with spectral density depending smoothly on the parameter
- (non-Gaussian) linear time series with coefficients depending smoothly on the parameter
- solutions to stochastic differential equations with drift or diffusion coefficient depending smoothly on the parameter, with asymptotics on the noise tending to zero, or the observation interval tending to infinity
- estimation of a tail index in extreme value theory
- counting processes with intensities depending smoothly on the parameter
- certain random fields
- hidden Markov models.

All these examples yield so-called "regular parametric models" in that the LAN conditions hold at every point in the parameter set and the dependence of the centering vectors and Fisher information on the parameter is continuous. In other examples it can be useful to utilize LAN *sub*models of larger, possibly infinite-dimensional models. For instance, in density estimation, inverse problems, nonparametric estimation and semiparametric modeling [e.g., Ibragimov and Has'minskii (1981), Begun, Hall, Huang and Wellner (1983), Koshevnik and Levit (1976), Millar (1979, 1983, 1985), Pfanzagl and Wefelmeyer (1982), Donoho and Liu (1991), Bickel, Klaassen, Ritov and Wellner (1993)].

Given the large domain of attraction of the normal distribution in the central limit theorem, the widespread occurrence of LAN is not surprising. One may ask if other types of approximations occur naturally.

Le Cam has studied the asymptotics for experiments based on independent observations in some detail [e.g., Le Cam (1969, 1974)]. Under the assumption that the individual observations are asymptotically negligible in a statistical sense, he derived a full characterization of the possible limit experiments, and criteria for convergence. Within this general context the LAN experiments converge to very special Gaussian exeriments. General limits can be characterized as *infinitely divisible experiments*, or more concretely as mixtures of Gaussian experiments (not necessarily linear in the parameter) and *Poisson experiments*. The latter consist of observing a Poisson process on an abstract space, with intensity measure depending on the parameter. Interesting simple examples of Poisson experiments arise for parametric models in which the density does not depend differentially on the parameter, the simplest and best known case being the experiment consisting of an i.i.d. sample from the uniform distribution, which is "asymptotically exponential" [see, e.g., Ibragimov and Has'minskii (1981) and Pflug (1983)].

As is clear from the preceding list of examples, local asymptotic normality is not limited to independent observations, but is a useful concept in many situations involving stochastic processes. In such situations LAN often arises from an application of the martingale central limit theorem to a two-term Taylor approximation to the log likelihood. However, a two-term quadratic expansion of the log likelihood process will not always give a sufficient approximation. Such an approximation is obviously impossible if the likelihood is not smooth in the parameter, but even for a very smooth likelihood a Taylor expansion may not do. One example is when the Fisher information is zero; a much less trivial example is the autoregressive process $X_t = \theta X_{t-1} + Z_t$ in the explosive case $\theta > 1$ with non-Gaussian innovations Z_t [see Koul and Pflug (1990)].

If approximation by a quadratic Taylor expansion is possible, a situation that has become formalized as *locally asymptotically quadratic* or LAQ, then the most frequently occurring non-LAN case is *local asymptotic mixed normality* or LAMN. The matrices J_θ in the quadratic term must then be taken random and dependent on n, and the sequence $\Delta_{n,\theta}$ converges in distribution to a Gaussian scale mixture. This type of situation is relatively well-known, both because it occurs frequently and because the limit experiment is easy to analyze [see, e.g., Jeganathan (1982, 1995), Basawa and Scott (1983) and Gushchin (1995)].

Thus, there are many possible limits for statistical experiments. Le Cam's later work moved away from the linear Gaussian approximations and addressed the approximation problem in general, even though he would develop further theory for the LAN situation throughout his career. For instance, Le Cam (1985a) addressed global approximations for (generalized) LAN experiments (see Section 9), and Le Cam and Yang (1988b) studied the conditions under which experiments consisting of observing transformations $T_n(X_n)$ of observations X_n from an LAN sequence of experiments are LAN.

3. Contiguity. In his 1960 paper on local asymptotic normality Le Cam also introduced the concept of contiguity. The exact definition as given there has been copied into many books, even though Le Cam (1986) prefers a definition in terms of the limits of binary experiments. The famous *first three lemmas*, although not stated in the form of separate lemmas, also appear in Le Cam (1960a). They became well known through their statement by Hájek and Šidák (1967), who used them effectively to compare the asymptotic power of rank tests.

The 1960 definition of contiguity says that two sequences of probability measures P_n and Q_n defined on measurable spaces $(\mathcal{X}_n, \mathcal{A}_n)$ are *contiguous* if for any sequence of events $A_n \in \mathcal{A}_n$ one has $P_n(A_n) \to 0$ if and only if $Q_n(A_n) \to 0$. This implies that the sequences of measures P_n and Q_n do not separate asymptotically: given data from P_n or Q_n it is impossible to tell with certainty from which of the two sequences the data is generated, at least in an asymptotic sense, as $n \to \infty$. Indeed, if P_n and Q_n are contiguous, and ϕ_n is a sequence of tests with error probabilities $P_n\phi_n$ for testing the null hypothesis

P_n satisfying $P_n\phi_n \to 0$, then the power $Q_n\phi_n$ at the alternative Q_n satisfies $Q_n\phi_n \to 0$ as well.

Actually, contiguity implies more: contiguity is "asymptotic absolute continuity," meaning that it is possible to derive asymptotic probabilities computed under P_n from those computed under Q_n. This is the content of Le Cam's third lemma, which we discuss below.

The interpretation of contiguity as "asymptotic absolute continuity" comes out clearer in Le Cam's (1986) definition. This says that the sequences P_n and Q_n are contiguous if every limit point (P, Q) of the experiments (P_n, Q_n), in the sense of convergence of experiments defined in Section 5, has that P and Q are absolutely continuous.

These definitions are simple enough. The genius is that the two types of definitions are equivalent, and that the concept of contiguity is so useful.

There are several different technical criteria to decide whether two sequences are contiguous. The most important one is in terms of the log likelihood ratios

$$\log \frac{dQ_n}{dP_n}.$$

Contiguity is equivalent to this sequence of log likelihood ratios being asymptotically tight in $\mathbb{R}$, both when computed under P_n and under Q_n. In the case that this sequence under P_n is asymptotically normal $N(\mu, \sigma^2)$, then contiguity is equivalent to $\mu = -\frac{1}{2}\sigma^2$. This surprising result is explained by the fact that absolute continuity of two probability measures P and Q is equivalent to $\mathrm{E}_P(dQ/dP) = 1$. If $\log(dQ/dP)$ is $N(\mu, \sigma^2)$-distributed under P, then this equation is valid if and only if $\mu = -\frac{1}{2}\sigma^2$.

The curious equation $\mu = -\frac{1}{2}\sigma^2$ arises naturally for locally asymptotically normal experiments, where it results from the expansion (2.1) of the log likelihood in linear and quadratic terms: the sequence $\log(dP_{n,\theta+\delta_n h_n}/dP_{n,\theta})$ is asymptotically normal with mean the quadratic term $-\frac{1}{2}h^T J_\theta h$ and variance $h^T J_\theta h$. Thus the sequences $P_{n,\theta+\delta_n h_n}$ and $P_{n,\theta}$ in a LAN experiment are contiguous for every bounded sequence h_n.

The characterization in terms of the log likelihood ratios suggests ways to create "contiguous alternatives" if a sequence P_n is given: define $dQ_n = h_n dP_n$, where $h_n \approx 1$ is a perturbation such that $\log h_n = \log(dQ_n/dP_n)$ behaves appropriately.

Contiguity has turned out to be a wonderful tool in many proofs, where one is given a choice to prove convergence in probability to zero under the measure of interest, or under any other convenient, contiguous sequence. (For a powerful use of this see Le Cam's proof of his Bernstein–von Mises theorem, mentioned in Section 12.) However, the application of contiguity that has made it popular is in the comparison of statistical tests. Here one is given a sequence of tests ϕ_n concerning a parameter h attached to a statistical model $(P_{n,h} : h \in H)$ and corresponding power functions

$$\pi_n(h) = P_{n,h}\phi_n.$$

If P_{n,h_0} and P_{n,h_1} are asymptotically separated, then any "good" sequence of tests of the null hypothesis h_0 versus the alternative h_1 will have $\pi_n(h_0) \to 0$ and $\pi_n(h_1) \to 1$. Such alternatives are not of much interest to compare the quality of two sequences of tests. On the other hand, contiguous alternatives will not allow this type of degeneracy, and hence may be used to pick a best test, or compute a *relative efficiency* of two given sequences of tests. Such contiguous alternatives may be given through the context, for instance of a parametric model. In particular, for LAN models the measures corresponding to parameters $\theta_{n,1}$ and $\theta_{n,2}$ at distance $\theta_{n,1} - \theta_{n,2} = O(\delta_n)$ are contiguous. Alternatively, contiguous alternatives may be constructed for the purpose of power comparisons through the perturbation method described previously.

To prevent asymptotic separation of alternative hypotheses P_n and Q_n the full force of contiguity is not needed. Contiguity has a further use, which is to alleviate the problem of computing the limiting distribution of a test statistic under a (contiguous) alternative. This technique is skillfully applied to rank procedures in Hájek and Šidák (1967), and has since become a standard tool in the asymptotic analysis of tests. The basic procedure, known as Le Cam's third lemma, can be found in Le Cam (1960a).

LEMMA 3.1 (Third lemma). *If P_n and Q_n are contiguous sequences of probability measures and T_n is a sequence of statistics such that $(T_n, dQ_n/dP_n)$ converges in distribution under P_n to a vector (T, V), then T_n converges under Q_n in distribution to the law L defined by $L(B) = \mathrm{E}\mathbb{1}_B(T)V$.*

The proof of this lemma requires some technical work, but the idea is simple. If P_n and Q_n are absolutely continuous, then

$$Q_n(T_n \in B) = \int \mathbb{1}_B(T_n)\, dQ_n = \int \mathbb{1}_B(T_n)\, \frac{dQ_n}{dP_n}\, dP_n.$$

By assumption the vector $(T_n, dQ_n/dP_n)$ is under P_n asymptotically distributed as (T, V). This suggests that the right-hand side is asymptotic to $\mathrm{E}\mathbb{1}_B(T)V$. Contiguity is necessary and sufficient to justify passing to the limit in this argument.

As we noted, in a LAN situation the sequence dQ_n/dP_n is log normally distributed. If also the sequence T_n is asymptotically normal, then typically the sequence of vectors $(T_n, \log(dQ_n/dP_n))$ is asymptotically multivariate normal. For ease of notation suppose that under P_n it is asymptotically distributed as $(T, \log V)$. It is an easy computation to see that in that case the distribution L in Le Cam's third lemma is a normal distribution with mean $\mathrm{E}T + \mathrm{cov}(T, \log V)$ and the same covariance matrix as T. Thus passing from a given distribution to a contiguous alternative typically has the result of shifting the mean of an asymptotically normal sequence of statistics, leaving the variance the same.

In the testing situation, with asymptotically normal test statistics T_n, it follows that a change of measure from a null hypothesis P_{n,h_0} to a contiguous alternative P_{n,h_1} induces a change of asymptotic mean in the test statistics T_n equal to the asymptotic covariance between T_n and $\log(dP_{n,h_1}/dP_{n,h_0})$ and no change of variance. It follows that good test statistics have a large (asymptotic) covariance with the log likelihood ratios.

As a particular example consider the power of linear rank tests. We follow Hájek (1962) and Hájek and Šidák [(1967), pages 210–216]. Let $R_{n,1}, \ldots, R_{n,n}$ be the ranks of a set of independent real-valued observations $X_1, \ldots, X_n$. To test the null hypothesis that $X_1, \ldots, X_n$ are an i.i.d. sample from an unknown density f, Hájek and Šidák propose the linear rank statistic

$$S_n = \sum_{i=1}^{n} (c_{n,i} - \bar{c}_n) a_n(R_{n,i}),$$

where a_n are "scores" (a map from $\{1, \ldots, n\}$ to $\mathbb{R}$), and the $c_{n,i}$ are given constants. To investigate which constants $c_{n,i}$ or scores are appropriate, we could imagine that we wish to test the null hypothesis versus some particular alternative hypothesis. Hájek and Šidák are interested in location problems and consider the null hypothesis that the vector $(X_1, \ldots, X_n)$ is sampled from the density

$$p_n(x_1, \ldots, x_n) = \prod_{i=1}^{n} f(x_i),$$

versus the alternative hypothesis, that, for given constants $d_{n,i}$ with $\bar{d}_n = 0$, they are sampled from

$$q_n(x_1, \ldots, x_n) = \prod_{i=1}^{n} f(x_i - d_{n,i}).$$

If the density f is appropriately differentiable and the constants $d_{n,i}$ are sufficiently regular [see conditions (4) and (5) of Hájek and Šidák], then, with I_f the Fisher information for location of f,

$$\Lambda_n := \log \frac{\prod_{i=1}^{n} f(X_i - d_{n,i})}{\prod_{i=1}^{n} f(X_i)} \approx \sum_{i=1}^{n} d_{n,i} \frac{f'}{f}(X_i) - \frac{1}{2} \sum_{i=1}^{n} d_{n,i}^2 I_f. \tag{3.1}$$

This expansion takes the form of an LAN expansion. If the constants $d_{n,i}$ satisfy the conditions as mentioned, then it follows that the sequences of laws P_n and Q_n of $(X_1, \ldots, X_n)$ under null and alternative hypotheses are contiguous. In an earlier chapter Hájek and Šidák had already obtained the asymptotic distribution of the

sequence of linear rank statistics under the null hypothesis, based on the expansion

$$S_n \approx \sum_{i=1}^{n} (c_{n,i} - \bar{c}_n)\phi\big(F(X_i)\big), \tag{3.2}$$

where ϕ is the "score generating function" for the scores $a_n(i)$. [Roughly $a_n(i) = \phi(i/n)$, but we omit the details.] Combined, the two approximations (3.1) and (3.2) readily yield the joint asymptotic normality of the pair (S_n, Λ_n) under the null hypothesis. An application of Le Cam's third lemma yields that under the alternative hypothesis the test statistic is asymptotically normal with the mean shifted by the covariance

$$\sum_{i=1}^{n} (c_{n,i} - \bar{c}_n) d_{n,i} \int \frac{f'}{f}(x)\phi\big(F(x)\big) f(x)\, dx.$$

To construct good test statistics for this particular alternative we need to maximize this expression relative to the asymptotic standard deviation of S_n.

This leads to such a conclusion as that the Wilcoxon statistics are asymptotically optimal test statistics if the unknown null density f belongs to the logistic family. With the limit distributions under alternatives in hand we can also easily compute the asymptotic relative efficiency of, for instance, the Wilcoxon test versus the t-test, for a variety of f.

4. Minimax and convolution theorems. Theorems that in some way show that a normal distribution with mean zero and covariance matrix the inverse of the Fisher information is a "best possible" limit distribution have a long history, starting with Fisher in the 1920s and with important contributions by Cramér, Rao, Stein, Rubin, Chernoff and others, which we shall not recall here. Of course, "the" theorem referred to is not true, at least not without a number of qualifications. Le Cam contributed in various ways to an understanding of this issue, and eventually gave a complete explanation.

Hájek (1970, 1972) formulated and proved two theorems, using different types of qualifications, which are now considered as most appropriate. For instance, his formulations, originally for parametric locally asymptotically normal models, have been the templates for similar results for nonparametric and semiparametric models in the 1980–1990s. The idea of studying a minimax risk derives from statistical decision theory, and the application of local minimaxity in asymptotics can be traced back to Le Cam (1953) and Chernoff (1956). Apparently Hájek's convolution theorem came to Le Cam as "a bolt out of the blue" (R. Beran, personal communication), even though it is now common to speak of the Hájek–Le Cam convolution theorem, as Le Cam proved major generalizations.

Both of Hájek's theorems consider a sequence of experiments $(P_{n,\theta} : \theta \in \Theta)$ that is LAN (2.1) at a given point θ. The point θ is fixed throughout.

THEOREM 4.1 (LAM). *For any estimator sequence T_n and bowl-shaped loss function $\ell : \mathbb{R}^k \to [0, \infty)$,*

$$\sup_c \liminf_{n\to\infty} \sup_{\|h\|<c, \theta+\delta_n h\in\Theta} \mathrm{E}_{\theta+\delta_n h} \ell\big(\delta_n^{-1}(T_n - \theta - \delta_n h)\big) \geq \int \ell \, dN(0, J_\theta^{-1}). \tag{4.1}$$

THEOREM 4.2 (Convolution). *Let T_n be sequence of estimators such that $\delta_n^{-1}(T_n - \theta - h\delta_n) \rightsquigarrow L_\theta$ under $P_{n,\theta+h\delta_n}$ for some fixed distribution L_θ and every $h \in \mathbb{R}^k$. Then there exists a probability measure M_θ such that $L_\theta = N(0, J_\theta^{-1}) * M_\theta$.*

Hájek's proofs of these results are long and technical, based on approximations as in Le Cam (1960a). A key ingredient is his Lemma 3.3 [Hájek (1972), page 185], which derives directly from Le Cam [(1960a), page 84, paragraph 5], and shows that the centering variables $\Delta_{n,\theta}$ in (2.1) are "distinguished." Le Cam realized that there is a simpler and more general approach, based on the notion of a *limit experiment*. As we noted the experiments $(P_{n,\theta+\delta_n h} : h \in \mathbb{R}^k)$ are asymptotically like the normal experiment $(N(h, J_\theta^{-1}) : h \in \mathbb{R}^k)$. The convolution and minimax theorems may be derived from viewing the normal experiment as a whole as a "lower bound" for the sequence of experiments $(P_{n,\theta+\delta_n h} : h \in \mathbb{R}^k)$. Le Cam (1972a) made this idea rigorous in great generality by introducing an appropriate concept of convergence of experiments. (See Section 5.) It was obvious from his earlier work that the sequence $(P_{n,\theta+\delta_n h} : h \in \mathbb{R}^k)$ converges to the normal experiment given previously, according to this notion.

A deeper understanding of Hájek's theorems can be obtained by relating the criterion used to judge a sequence of estimators T_n to the same criterion applied to estimators in the limit experiment $(N(h, J_\theta^{-1}) : h \in \mathbb{R}^k)$. In the case of the minimax theorem the local asymptotic minimax risk at θ, the infimum of the left side of (4.1) over all possible estimators sequence T_n, is lower bounded by the minimax risk in the limit experiment, given by

$$\inf_T \sup_{h\in\mathbb{R}^k} \mathrm{E}_h \ell(T - h), \tag{4.2}$$

the infimum being over all (randomized) estimators T based on an observation X from the $N(h, J_\theta^{-1})$ distribution. [The supremum over $h \in \mathbb{R}^k$ corresponds to θ being an inner point of the parameter set Θ, so that the local parameter sets $\delta_n^{-1}(\theta - \Theta)$ will grow to $\mathbb{R}^k$, as $n \to \infty$.] It is standard statistics to compute the minimax risk in a normal location experiment, at least for nice loss functions and a completely "unknown" location parameter, as in (4.2). For the symmetric, bowl-shaped loss functions used in Hájek's theorem the minimax estimator is $T(X) = X$, and the minimax risk is the right-hand side of (4.1), as $X - h$ is $N(0, J_\theta^{-1})$-distributed under every h, so that

$$\sup_h \mathrm{E}_h \ell(X - h) = \int \ell \, dN(0, J_\theta^{-1}).$$

An added insight here is that the minimax risk (4.2) is a valid lower bound for the local asymptotic minimax risk for any loss function that "survives" taking limits. This includes all lower semi-continuous functions $\ell : \mathbb{R}^k \to [0, \infty)$ with the property that $\liminf_{\|y\| \to \infty} \ell(y) \geq \ell(x)$ for every $x \in \mathbb{R}^k$. Similarly, the minimax theorem remains valid at boundary points θ of the parameter set, provided the sequence of local parameter sets $\delta_n^{-1}(\theta - \Theta)$ converges in a suitable sense to a limit H and the supremum in (4.2) is restricted to $h \in H$. However, it may not be possible to evaluate (4.2) explicitly in these cases.

Hájek's convolution theorem puts a regularity restriction on the estimator sequence. A sequence of estimators T_n is *regular* at θ if the sequence $\delta_n^{-1}(T_n - \theta - \delta_n h)$ converges under $\theta + \delta_n h$ to a fixed limit distribution, independent of h, for every h. Le Cam realized that the corresponding regularity restriction on estimators T in the limit experiment is location equivariance. We can see this by matching up the law of $\delta_n^{-1}(T_n - \theta) - h$ under $P_{n,\theta+\delta_n h}$ and the law of $T - h$ under $N(h, J_\theta^{-1})$. Regularity requires that the first law tends to L_θ as $n \to \infty$, and hence for a matching estimator T the second must be L_θ as well, for every h. Thus the distribution under h of $T - h$ is independent of h. Such estimators T that are "equivariant-in-law" are rare. The identity $T(X) = X$ is an example. It had already been proved by Boll (1955) that the distribution of a general equivariant estimator could be written as a convolution: they can be represented as the sum $X + U$ of the "optimal estimator" X and random, ancillary noise U independent of X.

The preceding exposition is written in readily interpretable statistical language. For flavor, we include a formulation of the convolution theorem by Le Cam. It is concrete in that it concerns the full shift case, albeit on a general locally compact group Θ [Le Cam (1986), page 128]:

> The operator S^α in (ii) below maps a given distribution F into the distribution $B \mapsto F(\alpha^{-1}B)$. Condition (iii) that the sequence X_n is *distinguished* means that its set of limit distributions $\{F_\theta : \theta \in \Theta\}$ in (i) are a limit experiment for the sequence of experiments $\mathcal{F}_n$.
>
> The above arguments lead almost immediately to a nice result of J. Hájek. A form of it is as follows. Consider the case where Θ and Z are one and the same locally compact group G. For each n let $\mathcal{F}_n = \{P_{\theta,n} : \theta \in \Theta\}$ be an experiment indexed by Θ. Consider also two statistics, say X_n and Y_n, available on $\mathcal{F}_n$ and taking values in $G = Z = \Theta$.
>
> PROPOSITION 2. Assume that the following conditions are satisfied:
>
> (i) The distributions $F_{\theta,n} = \mathcal{L}(X_n \mid P_{\theta,n})$ and $G_{\theta,n} = \mathcal{L}(Y_n \mid P_{\theta,n})$ converge, respectively, to limits F_θ and G_θ.
>
> (ii) The limits F_θ and G_θ are such that $F_{\alpha\theta} = S^\alpha F_\theta$ and $G_{\alpha\theta} = S^\alpha G_\theta$.
>
> (iii) For the sequence of experiments $\{\mathcal{F}_n\}$ the X_n are a distinguished sequence.
>
> (iv) The F_θ are absolutely continuous with respect to the Haar measure of G. Furthermore, G admits almost invariant means.
>
> Then there is a probability measure Q such that $G_\theta = F_\theta * Q$ for all θ.

5. Limits of experiments. The observations concerning limit experiments and lower bounds were laid down in Le Cam (1972a). Le Cam (1972a) opens with:

> In a recent paper J. Hájek (1970) proved a remarkably simple result on the limiting distribution of estimates of a vector parameter θ. It turns out that this result, as well as many of the usual statements about asymptotic behavior of tests or estimates, can be obtained by a general procedure which consists roughly in passing to the limit first and then arguing the case for the limiting problem. This passage to the limit relies on some general facts which perhaps are not entirely elementary. They depend heavily on techniques of L. Le Cam (1964a). However, these general facts are of interest by themselves. If they are taken for granted the basic result of Hájek (1970) and many results of A. Wald (1943) become available immediately.

The paper Le Cam (1972a) appeared in the Proceedings of the Sixth Berkeley Symposium, where Hájek (1972) presented his local asymptotical minimax theorem. Le Cam gives explicit credit to Hájek for helping shape the paper [Le Cam (1972a), page 246]:

> Any resemblance between our results and those of Hájek is not entirely accidental, since the present paper was greatly modified after Hájek's presentation during the Symposium.

The first line of this remark, made in the introduction of Le Cam (1972a), should probably be understood as an expression of appreciation for Hájek's achievement. It is not very accurate, because the overlap between the papers is small.

As he acknowledged in the preceding quotation, Le Cam's (1972a) paper is far from elementary. The general facts from his 1964 paper "which perhaps are not entirely elementary" concern convergence of experiments in a "deficiency measure," which we review in more detail in Section 7. For now we only describe the meaning of zero deficiency.

The deficiency $\delta(\mathcal{E}, \mathcal{F})$ of an experiment $\mathcal{E}$ relative to another experiment $\mathcal{F}$, with the same parameter set but possibly a different sample space, is zero if and only if statistical aims achievable in the experiment $\mathcal{F}$ can be achieved at least as well in the experiment $\mathcal{E}$. The experiment $\mathcal{F}$ is then also said to be *weaker* than $\mathcal{E}$. This notion can be made precise in several equivalent ways, by using risk functions, randomizations, or through the matching-in-law of statistics. The third characterization is most relevant to the implications of limits of experiments and says that given a statistic in the experiment $\mathcal{F}$, one can find a (randomized) statistic in $\mathcal{E}$ with exactly the same set of laws.

A concrete statement of the third type is as follows. (It is a theorem or a definition, depending on one's starting point.) We define a *randomized estimator* T in an experiment with sample space $(\mathcal{X}, \mathcal{A})$ and values in a metric space $\mathbb{D}$ as a (Borel) measurable map $T : \mathcal{X} \times [0, 1] \to \mathbb{D}$. We evaluate its "law under P" as the law of $T(X, U)$ for X having law P and U being a uniform variable independent of X, that is, as $(P \times \lambda) \circ T^{-1}$, for λ the Lebesgue measure on $[0, 1]$.

THEOREM 5.1. *Let* $\mathcal{E} = (\mathcal{X}, \mathcal{A}, P_h : h \in H)$ *and* $\mathcal{F} = (\mathcal{Y}, \mathcal{B}, Q_h : h \in H)$ *be dominated experiments. Then* $\delta(\mathcal{E}, \mathcal{F}) = 0$ *if and only if for every Polish space* $\mathbb{D}$ *and every randomized estimator* T *in* $\mathcal{F}$ *with values in* $\mathbb{D}$ *there exists a randomized estimator* S *in* $\mathcal{E}$ *such that* $(P_h \times \lambda) \circ S^{-1} = (Q_h \times \lambda) \circ T^{-1}$ *for every* $h \in H$.

Since we can measure statistical difficulty through the possible sets of laws of statistics, we should prefer $\mathcal{E}$ over $\mathcal{F}$.

The assumptions of domination and Polishness can be removed at the expense of technical difficulties. In Section 8 we discuss how Le Cam solved these difficulties.

A main result of Le Cam (1972a) can be stated as follows.

THEOREM 5.2. *Let* $S_n : \mathcal{X} \to \mathbb{D}$ *be a sequence of statistics defined in experiments* $\mathcal{E}_n = (\mathcal{X}_n, \mathcal{A}_n, P_{n,h} : h \in H)$ *and with values in a fixed metric space* $\mathbb{D}$ *such that* $S_n \rightsquigarrow Q_h$ *under* $P_{n,h}$ *for every* $h \in H$. *Set* $\mathcal{F} = (Q_h : h \in H)$. *Then* $\delta(\mathcal{E}, \mathcal{F}) = 0$ *for every weak limit point* $\mathcal{E}$ *of the sequence* $\mathcal{E}_n$.

The theorem relates two types of weak convergence. The first type is the usual convergence in distribution $\rightsquigarrow$, here applied to statistics S_n and yielding a collection $\mathcal{F} = (Q_h : h \in H)$ of limit laws of the sequence S_n (on the Borel σ-field of $\mathbb{D}$). The second type concerns weak convergence of a sequence of statistical experiments, a type of convergence which we still need to define. It is applied here to subsequences or subnets of the sequence of experiments $\mathcal{E}_n$. The two types of convergence concern different types of objects (statistics and experiments) and hence cannot be compared directly. However, the theorem says that the usual weak convergence leads to experiments that are weaker than the experiments obtained as limits in the other type of convergence.

As we noted there are several ways to interpret zero deficiency between experiments. Using the interpretation through the existence of (randomized) estimators, we can reformulate the theorem as follows.

THEOREM 5.3. *Let* $S_n : \mathcal{X} \to \mathbb{D}$ *be a sequence of statistics defined in experiments* $\mathcal{E}_n = (\mathcal{X}_n, \mathcal{A}_n, P_{n,h} : h \in H)$ *and with values in a fixed Polish metric space* $\mathbb{D}$ *such that* $S_n \rightsquigarrow Q_h$ *under* $P_{n,h}$ *for every* $h \in H$. *Assume that the sequence* $\mathcal{E}_n$ *converges to* $\mathcal{E} = (\mathcal{X}, \mathcal{A}, P_h : h \in H)$. *If* $\mathcal{E}$ *is dominated, then there exists in* $\mathcal{E}$ *a randomized estimator* $T : \mathcal{X} \times [0, 1] \to \mathbb{D}$ *such that* T *possesses law* Q_h *under* P_h, *for every* $h \in H$.

Before giving a definition of convergence of experiments, let us see how the theorem can be used to prove Hájek's theorems. We need the information that the sequence of local experiments $\mathcal{E}_n := (P_{n,\theta+\delta_n h} : h \in \mathbb{R}^k)$ extracted from an LAN sequence $(P_{n,\theta} : \theta \in \Theta)$ converges to the experiment $\mathcal{E} := (N(h, J_\theta^{-1}) : h \in \mathbb{R}^k)$ encountered previously. Next we consider a sequence of $\mathbb{R}^k$-valued statistics T_n

in the experiments $(P_{n,\theta} : \theta \in \Theta)$ and assume that $S_n := \delta_n^{-1}(T_n - \theta)$ converges in distribution under $P_{n,\theta+\delta_n h}$ to a limit distribution Q_h. (The parameter θ is fixed in this argument; we have suppressed it from the notation.) By the preceding theorem there exists a (randomized) statistic $T : \mathbb{R}^k \to \mathbb{R}^k$ in the Gaussian experiment $\mathcal{E}$ such that T possesses the law Q_h under h, for every $h \in \mathbb{R}^k$. In other words, the sequence $\delta_n^{-1}(T_n - \theta - \delta_n h)$ converges in distribution under $P_{n,\theta+\delta_n h}$ to the distribution of $T - h$ under h.

If the sequence of estimators T_n is regular with limit law L_θ, then it follows immediately that $T - h$ possesses law L_θ under every h. In group-statistics language the statistic T must be equivariant-in-law. We can now apply Boll's (1955) result to see that the distribution L_θ can be decomposed as in the convolution theorem.

We can similarly prove Hájek's LAM theorem. General weak convergence theory and lower semicontinuity of ℓ give that

$$\liminf_{n\to\infty} \mathrm{E}_{\theta+\delta_n h} \ell\big(\delta_n^{-1}(T_n - \theta - h\delta_n)\big) \geq \mathrm{E}_h \ell(T - h).$$

This is true for every local parameter h. We can now add first the $\sup_{\|h\|<c}$ within the lim inf on the left-hand side, making this bigger, and next the $\sup_c$ on the far left-hand side, making this still bigger, thus creating the left-hand side of (4.1). We may make the same changes on the right side, and keep the inequality, yielding a right-hand side

$$\sup_{h\in\mathbb{R}^k} \mathrm{E}_h \ell(T - h).$$

It is standard decision theory that this expression is minimized by $T(X, U) = X$, for any bowl-shaped loss function. For this estimator the expression in the display reduces to the right-hand side of (4.1).

This argument is based on the assumption that the sequence $\delta_n^{-1}(T_n - \theta - \delta_n h)$ converges in distribution under $P_{n,\theta+\delta_n h}$ to a limit distribution, for every h, but Hájek's minimax theorem is valid for every estimator sequence T_n. We can extend the preceding proof by a compactification argument. For instance, if we consider the estimators T_n to be maps in a compactification of $\mathbb{R}^k$, then the sequence T_n is automatically tight and will converge in distribution along subsequences, by Prohorov's theorem.

Le Cam would compactify in a different way, by interpreting Theorem 5.2 in a more abstract setting, involving "procedures" rather than statistics, an abstract definition of "statistical experiments" and an extended definition of "risk functions." Theorem 5.3 is a special, but statistically more transparent form of Le Cam's (1972a) main result, which is closer to Theorem 5.2. A direct proof of this special theorem is not necessarily easier than the proof of the corresponding abstract theorem, because it asserts the existence of concrete statistical objects: randomized estimators. In contrast, Theorem 5.2 yields "procedures" or "transitions." Only if one takes certain representation theorems

for linear functionals and operators for granted, the special theorem can become a corollary of the more abstract formulation. These are among the "general facts which perhaps are not entirely elementary," which Le Cam asked to take for granted at the beginning of his 1972 paper. (See the quotation at the beginning of this section.)

This difference is comparable to proofs of Prohorov's theorem in weak convergence theory. Prohorov's theorem (which was independently proved by Le Cam) says that every uniformly tight sequence of random elements X_n in a metric space possesses a weakly converging subsequence. It is relatively easy to extract a subsequence n' such that $\mathrm{E} f(X_{n'}) \to Lf$ for every bounded, continuous real function f and some positive linear map L, but it requires some effort to show that L is representable by a probability measure. The uniform tightness of the sequence X_n is most crucial in this second part of the argument, because it implies that the limit L must be a special type of functional.

Because the convergence of the local experiments of an LAN sequence of experiments to a Gaussian shift experiment in the sense of Le Cam (1972a) was clear from Le Cam's previous work [e.g., Le Cam (1969)], Le Cam could recover Hájek's theorems by analyzing the Gaussian limit experiment. Le Cam (1972a) recovers the convolution theorem within the context of general (i.e., possibly non-Gaussian) Euclidean shift experiments as limits, but pays relatively little attention to the minimax theorem. Le Cam (1979a) completed the analysis of Hájek's theorems by proving that the difference between two sequences of estimators that converge to a randomized estimator that is uniquely determined by its risk function must necessarily converge to zero. This applies for instance to LAM sequences of estimators of a parameter of dimension lower than three.

The great contribution of Le Cam (1972a) was to describe limiting experiments in general: one can apply the same arguments as given also in many other situations, including the LAMN or LAQ experiments mentioned before, the Poisson experiments arising when rescaling nonsmooth parametric experiments, or nonlinear Gaussian experiments arising as subexperiments in non- and semiparametric statistics. The "passing to the limit" will be the same each time, but analyzing the limit experiment may not give such nice and simple results as in the Gaussian case. For instance, one may be left with a statement that the asymptotic minimax risk is bounded below by the minimax risk in an experiment involving the observation of several Poisson processes with intensities depending on the parameter.

In that sense the assertion of the convolution theorem is special: even if one can define a notion of regularity in many settings, equivariant-in-law estimators are not always characterized by a convolution property.

The definition of weak convergence of experiments given in Le Cam (1972a) employs the deficiency distance of Le Cam (1964a): a sequence (or net) of experiments $\mathcal{E}_n = (P_{n,h} : h \in H)$ converges to a limit experiment $\mathcal{E} = (P_h : h \in H)$ if for all finite subsets $I \subset H$ $\Delta(\mathcal{E}_n^I, \mathcal{E}^I) \to 0$, where $\mathcal{E}_n^I = (P_{n,h} : h \in I)$ is the subexperiment with parameter restricted to I, and Δ is the deficiency distance,

as discussed in Section 7. This definition is statistically meaningful, but not the easiest definition to work with in examples. In most cases convergence of experiments is proved through marginal weak convergence of the log likelihood ratio processes. An alternative definition of convergence of $\mathcal{E}_n = (P_{n,h} : h \in H)$ to $\mathcal{E} = (P_h : h \in H)$ is that, for every $h_0 \in H$ and finite set $I \subset H$,

$$\left(\frac{dP_{n,h}}{dP_{n,h_0}}\right)_{h\in I} \rightsquigarrow \left(\frac{dP_h}{dP_{h_0}}\right)_{h\in I}, \tag{5.1}$$

where the laws of the vectors on the left are computed under P_{n,h_0} and the law of the vector on the right is computed under P_{h_0}. The wiggly arrow denotes ordinary weak convergence in $\mathbb{R}^I$.

It is not difficult to verify that the likelihood ratio process of an LAN sequence of experiments converges to the likelihood ratio process of a Gaussian experiment in this sense. Indeed, the likelihood ratios of the Gaussian experiment $(N(J_\theta h, J_\theta) : h \in R^k)$ are given by

$$\frac{dN(J_\theta h, J_\theta)}{dN(0, J_\theta)}(\Delta) = e^{h^T\Delta - (1/2)h^T J_\theta h}. \tag{5.2}$$

This Gaussian experiment corresponds to observing Δ with a $N(J_\theta h, J_\theta)$ distribution, and is, for fixed θ, equivalent to the experiment consisting of observing $J_\theta^{-1}\Delta$, which possesses a $N(h, J_\theta^{-1})$ distribution. The LAN condition (2.1) shows that the likelihood ratio process $dP_{n,\theta+\delta_n h}/dP_{n,\theta}$ possesses exactly the same form, up to a term that converges to zero in probability. This verifies (5.1) for $h_0 = 0$; verification for other values of h_0 is similar.

It is the realization of the equivalence of the two definitions, together with the theorem as stated previously, that create the breakthrough of Le Cam's (1972a) paper. Le Cam (1972a) did not emphasize the equivalence, but did make the connection between weak convergence in terms of the deficiency distance and the weak convergence of likelihood ratio processes (5.1) in the special case that H is finite and P_{n,h_0} is replaced by the sum $\mu_n = \sum_h P_{n,h}$. The laws of these likelihood ratio processes under μ_n are known as the *canonical* or *conical measures* of the experiments. This connection had previously been explored in Le Cam [(1969), page 14, Théorème 1, apparently found earlier, but unpublished] and Torgersen (1968, 1970).

6. Superefficiency. Le Cam has contributed to an understanding of the super efficiency phenomenon at various points in his career, using the new insights obtained to give sharper, prettier, or deeper results at each point. We have written a longer review of these contributions for Le Cam's Festschrift [Pollard, Torgersen and Yang (1997)] and therefore shall be briefer here. The result must be mentioned, because it is both incredibly pretty and extremely relevant.

In the 1940s an estimator sequence was said to be *superefficient* if its asymptotic variance was smaller than the inverse Fisher information as given by the Cramér–Rao bound for the variance of unbiased estimators. The existence of such

superefficient estimators appeared to indicate that, in the LAN case, the normal distribution $N(0, J_\theta^{-1})$ does not give a lower bound for asymptotic estimation.

Some of the superefficient estimators can be easily discarded, because their risk functions behave erratically. This includes for instance the Hodges estimator $T_n = \bar{X}_n \mathbb{1}_{|\bar{X}_n| < n^{-1/4}}$ of a normal mean based on a sample $X_1, \ldots, X_n$ from the $N(\theta, 1)$-distribution. This estimator is designed to have small risk at the parameter value $\theta = 0$, but as a result has, for any fixed n, large risk at other parameter values close to 0, even though the risk at every fixed parameter value decreases below J_θ^{-1} eventually, as $n \to \infty$. Not all superefficient estimators are bad, however. In particular, *shrinkage estimators*, discovered by Stein (1956) shortly after Le Cam (1953) wrote on superefficiency, are superefficient and are good estimators. Depending on the loss function they beat the usual estimators, which have a $N(0, J_\theta^{-1})$ limit distribution.

There are several ways to save the Cramér–Rao bound. One is to consider local maximum risk, and this culminated in Hájek's LAM theorem considered before. Hodges estimator fails the quality test imposed by Hájek's LAM theorem, because its LAM risk at $\theta = 0$ is infinite. However, a disadvantage of taking a maximum risk is that it may hide differences. For instance, in terms of the LAM risk the shrinkage estimators (which are really better in dimensions 3 and up) are equivalent to the usual estimators. This is because the risk functions are asymptotic to each other as the parameter tends to infinity, even though one is always below the other.

Another way to save the Cramér–Rao bound is to note that asymptotic superefficiency can occur only on very small sets of parameters, for instance null sets for the Lebesgue measure. Le Cam proved this for the first time in 1953, in his thesis. The following is a much nicer result, discovered by Le Cam later on. It can be deduced from results in Le Cam (1973b).

THEOREM 6.1. *Assume that the sequence of experiments* $(P_{n,\theta} : \theta \in \Theta \subset \mathbb{R}^k)$ *is LAN at every* θ *with norming rate* δ_n *and nonsingular Fisher information matrices* J_θ. *Let the maps* $\theta \to P_{n,\theta}$ *be measurable. Let* T_n *be an estimator sequence such that* $\delta_n^{-1}(T_n - \theta)$ *converges to a limit distribution* L_θ *under every* θ. *Then there exist probability distributions* M_θ *such that, for Lebesgue almost every* θ,

$$L_\theta = N(0, J_\theta^{-1}) * M_\theta.$$

This remarkable theorem yields the assertion of Hájek's convolution theorem at almost every parameter value θ, without having to impose the regularity requirement on the estimator sequence. The convolution property implies that the covariance matrix of L_θ, if it exists, must be bounded below by the inverse Fisher information, and much more. Such consequences were also obtained by direct arguments by other authors, for example, Bahadur (1964).

The reason that the theorem is true is even more remarkable and is the main focus of Le Cam (1973b): it is roughly true that any estimator sequence T_n is "almost Hájek regular" at almost every parameter θ, at least along a subsequence of $\{n\}$. Le Cam [(1986), pages 128–129] explains this as follows. Let $G_n(\theta, h)$ be the distribution of $\delta_n^{-1}(T_n - \theta)$ under $P_{n,\theta+\delta_n h}$ and define "shift operators" S_i^α working on these distributions by

$$G_n(\theta, h)S_1^\alpha = G_n(\theta + \alpha, h),$$
$$G_n(\theta, h)S_2^\alpha = G_n(\theta, h + \alpha).$$

Furthermore, let $S^\alpha G = G * \delta_\alpha$ denote the measure G shifted by the vector α. This notation implies the algebraic identity

$$G_n(\theta, h)S_1^{\delta_n\alpha} = S^{-\alpha}G_n(\theta, h)S_2^\alpha.$$

Le Cam [(1986), page 133] proceeds:

> On the right side the coefficient δ_n does not appear explicitly. On the left side the shift operates only by an amount $\delta_n\alpha$ which tends to zero. *Thus, passing to the limit one should expect that,* with a modicum of continuity, *one will obtain an invariance relation such as* $G = S^{-\alpha}GS_2^\alpha$. This is precisely what happens, but it is perhaps surprising, or even frustrating that the amount of continuity needed to obtain such a relation is very little indeed.

Next it turns out that one can get already very far, for instance as far as the preceding theorem, under just some measurability conditions, rather than continuity. This brings out both the beauty of the limit results, and their dangers. The example of shrinkage estimators shows that the asymptotic null sets are not necessarily small for finite n.

As this intuitive argument shows, this principle of automatic equivariance has nothing to do with LAN or Gaussian experiments. It is a consequence of rescaling a given Euclidean experiment. A concrete way to illustrate this is the following lemma, which shows that every estimator sequence in arbitrary experiments is almost regular in the sense of Hájek at almost every parameter.

LEMMA 6.2. *Let T_n be estimators in experiments $(P_{n,\theta} : \theta \in \Theta)$ indexed by a measurable subset Θ of $\mathbb{R}^k$. Assume that the map $\theta \to P_{n,\theta}(A)$ is measurable for every measurable set A and every n, and suppose that there exist distributions L_θ such that for Lebesgue almost every θ the sequence $r_n(T_n - \theta)$ tends under θ in distribution to a limit L_θ. Then for every $\gamma_n \to 0$ there exists a subsequence of $\{n\}$ such that, for Lebesgue almost every (θ, h), along the subsequence, the sequence $r_n(T_n - \theta - \gamma_n h)$ converges in distribution under $\theta + \gamma_n h$ to L_θ.*

As we saw, the Hájek regularity translates into equivariance of the matching estimator in the limit experiment. In the LAN case this next implies a convolution property. In other cases the restriction of equivariance may imply different

properties, but in most cases it will imply a lot. The form of the implication depends on the form of the limit experiment corresponding to the sequence $(P_{n,\theta+\delta_n h} : h \in \mathbb{R}^k)$. The experiment $(N(h, J_\theta^{-1}) : h \in \mathbb{R}^k)$, obtained under LAN, is a *full shift* of the $N(0, J_\theta^{-1})$ distribution under the additive group. If we replace the normal distribution by another Lebesgue absolutely continuous one, but retain the full shift property, then we shall again obtain a convolution theorem, but with a non-Gaussian kernel. This follows from Boll's (1955) result, who proved that the invariant law of an equivariant estimator in a dominated location experiment $(P \circ \delta_h : \in \mathbb{R}^k)$ always possesses the base distribution P as a convolution factor.

Shift experiments are common as limits, but they are not always full. An example of a partial shift is obtained by letting the nth experiment consist of observation of a sample of size n from the uniform measure on $[-\theta, \theta]$. This gives a shift of the distribution of a pair of two independent exponential variables over the diagonal in $\mathbb{R}^2$. (The intuitive explanation is that the two sufficient statistics for this experiment, the minimum and the maximum of the observations, are asymptotically exponential.) An example of a conditional shift experiment is obtained under LAMN. There is a great variety of other interesting limit experiments.

To each of these corresponds a superefficiency result. We give one simple example to illustrate this.

LEMMA 6.3. *Let T_n be estimators based on an i.i.d. sample of size n from the uniform distribution on $[0, \theta]$ such that the sequence $n(T_n - \theta)$ converges under θ in distribution to a limit L_θ, for every $\theta > 0$. Then $\int |x|^2 \, dL_\theta(x) \geq \theta^2$ for Lebesgue almost every θ.*

The final statement on the superefficiency problem was made by Le Cam (1973b). The treatment there, copied in Le Cam [(1986), pages 132–144] is complicated, because Le Cam avoids assuming that the estimators of interest, or the experiments in which they are defined, converge to limits, as we did in the preceding. Instead he imposes a weak topology for which one always has limit points and next shows that every limit point has the invariance property. This invariance (or equivariance) is valid both for sets of limit distributions and limit experiments. It can be deduced from this, that if the local experiments $(P_{n,\theta+\delta_n h} : h \in \mathbb{R}^k)$ converge to limit experiments, for every θ, then these limits must be "invariant," for almost every θ. Thus one can (almost) always apply group invariance arguments to derive optimal equivariant estimators, and hence a best possible limit distribution for a regular estimator sequence. The "invariance" takes the following form [simplified from Le Cam (1973b)].

THEOREM 6.4. *Suppose that the experiments $(P_{n,\theta+\delta_n h} : h \in \mathbb{R}^k)$ converge to limit experiments $\mathcal{E}_\theta = (P_{\theta,h} : h \in \mathbb{R}^k)$ such that the maps $h \mapsto P_{\theta,h}$ are*

continuous, for every θ. *Then for almost every* θ *there exists a measurable space* $(\mathcal{X}_\theta, \mathcal{A}_\theta)$ *and measurable bijections* $S_{\theta,h}: \mathcal{X}_\theta \to \mathcal{X}_\theta$ *such that* $S_{\theta,h_1} S_{\theta,h_2} = S_{\theta,h_1+h_2}$ *and such that* $\mathcal{E}_\theta$ *is equivalent to the experiment* $(\mathcal{X}_\theta, \mathcal{A}_\theta, P_\theta \circ S_{\theta,h}^{-1}: h \in \mathbb{R}^k)$.

Regarding the last sentence, any experiment that is equivalent to a given limit experiment, in the sense of having deficiency distance zero, is also a limit experiment. Thus a limit experiment is far from unique. The theorem asserts that it can be chosen to resemble a shift experiment, relative to the action of certain bijections. The example of partial shifts shows that one can not always arrange it so that $\mathcal{X}_\theta = \mathbb{R}^k$ and $S_{\theta,h}$ is a translation by h.

7. Deficiency. Le Cam introduced his deficiency distance between two experiments in 1964. The motivation came from two sources. We quote from Le Cam's review of Torgersen's book [Le Cam (1992a)]:

> Statisticians have been "comparing experiments" for a long time. One can see this just by glancing at R. A. Fisher's The Design of Experiments (1935) and at Neyman's "On two aspects of the representative method" (1934). However, in most instances, the comparisons were in terms of the performance of some special test or estimation procedure. The subject of the volume under review is said to have started in 1949 by a suggestion of J. von Neumann, quickly followed by a Rand Memorandum of Bohnenblust, Shapley and Sherman [(1949), unpublished].
>
> David Blackwell and Charles Stein soon recognized the statistical nature of the Rand Memorandum. Within the next years they proved, under a variety of restrictions, one of the main theorems of the subject. The result is often called the Blackwell–Sherman–Stein theorem. One of the visible differences with previous work of, say, J. Neyman, is that Blackwell and Stein looked at *all* the possible loss functions and risk functions for *all* the decision problems for the solution of which the experiment might have been performed.
>
> It was a great success of the early authors (Blackwell, Sherman, Stein) and of later ones such as V. Strassen, to show that "$\mathcal{E}$ better than $\mathcal{F}$" means that there is a Markov kernel T that applied to $\mathcal{X}$ and the P_θ reproduces measures Q_θ on $\mathcal{B}$. In symbols $Q_\theta = T P_\theta$. That is $\mathcal{F}$ is reproducible from $\mathcal{E}$ by "tossing coins."
>
> Another current of ideas came from the asymptotic statistical theory where one wanted to approximate some experiments by simpler ones. This can be said to have started, in a very special case, by a paper of A. Wald (1943). Eventually, the author of this review introduced the ideas of "deficiencies," "distance," weak convergence and the like.

Le Cam's 1964 paper can be seen as an attempt to make the Blackwell–Sherman–Stein quantitative, thus enabling him to address the problem of approximation of experiments initiated by Wald. The Blackwell–Sherman–Stein theorem is a statement that one statistical experiment $\mathcal{E}$ is "better" or "more informative" than another experiment $\mathcal{F}$ if and only if the second experiment can be obtained from the first through "randomization." In the terminology of Le Cam (1964a) this

means that the deficiency $\delta(\mathcal{E}, \mathcal{F})$ of $\mathcal{E}$ relative to $\mathcal{F}$ is zero. The major advance of Le Cam (1964a) was to quantify "deficiency" through a number $\delta(\mathcal{E}, \mathcal{F})$, thus enabling him to put asymptotic approximation and comparison of experiments within one framework.

Le Cam's 1964 paper was actually written in the late 1950s. Le Cam liked telling the story of how and why publication of the paper was delayed. From his own perspective he had written the paper as mathematically straightforward as possible, thus making the subject simple, but the referees did not recognize that kind of simplicity, and questioned its relevance to statistics. It is in this paper that the abstract functional-analytic machinery that Le Cam would use in most of his later papers first appears. The paper is very similar in spirit to the first chapters of his 1986 book.

Let us first look at the statistical content of deficiency. According to Wald's decision theory two statistical experiments can be compared by the set of risk functions available in each of them. In Wald's theory one is given an experiment $(\mathcal{X}, \mathcal{A}, P_\theta : \theta \in \Theta)$, a decision space $(\mathbb{D}, \mathcal{D})$ and a loss function $\ell : \Theta \times \mathbb{D} \to \mathbb{R}$. A decision procedure is a Markov kernel $(x, D) \mapsto \tau_x(D)$ from $(\mathcal{X}, \mathcal{A})$ into $(\mathbb{D}, \mathcal{D})$, with the interpretation that given an observation $x \in \mathcal{X}$ one chooses a decision from $\mathbb{D}$ according to the distribution τ_x. The risk of this procedure is defined to be the function

$$\theta \mapsto R(\theta; \tau) := \int_{\mathcal{X}} \int_{\mathbb{D}} \ell(\theta, y) \tau_x(dy)\, dP_\theta(x).$$

The purpose in the Wald framework is to find statistical procedures τ with a small risk function. Thus it is natural to say that the experiment $\mathcal{E}$ is "more informative" or "better" than the experiment $\mathcal{F}$ if for every procedure τ in $\mathcal{F}$ there is a procedure σ in $\mathcal{E}$ with $R(\theta; \sigma) \le R(\theta; \tau)$ for every $\theta \in \Theta$. The realization of Blackwell–Sherman–Stein was that this is equivalent to the existence of a *randomization* of $\mathcal{E}$ that exactly produces $\mathcal{F}$.

The idea is the same as in a proof that a sufficient statistic indeed contains all information on a parameter. In that example we wish to prove that observing only the sufficient statistic is as informative as observing the original observation. The randomization is here the conditional law of the observation given the sufficient statistic. Because it is assumed to be free of the parameter, this conditional law can be used to generate a (pseudo) observation given the observed value of the sufficient statistic. This pseudo observation is just as good as a real observation, because it has the same laws.

A general randomization within the Wald setup is a Markov kernel T from the sample space of $\mathcal{E} = (\mathcal{X}, \mathcal{A}, P_\theta : \theta \in \Theta)$ into the sample space of $\mathcal{F} = (\mathcal{Y}, \mathcal{B}, Q_\theta : \theta \in \Theta)$. Given an observation x in $\mathcal{E}$ we can generate an "observation" y in the sample space of $\mathcal{F}$ according to the Markov kernel T_x. If x is sampled from P, write TP for the distribution of y, that is,

$$TP(B) := \int T_x(B)\, dP(x). \tag{7.1}$$

If $TP_\theta = Q_\theta$ for every $\theta \in \Theta$ and x is generated from $P_\theta \in \mathcal{E}$, then the corresponding y will be generated from $Q_\theta \in \mathcal{F}$ and hence be exactly as an observation in $\mathcal{F}$. If this is true, then we can produce an observation in $\mathcal{F}$ from an observation in $\mathcal{E}$, and hence the experiment $\mathcal{E}$ is better in the sense of comparison of the available loss functions, because we can match any procedure in $\mathcal{F}$ by a procedure in $\mathcal{E}$: first produce y from x, next proceed with y as in $\mathcal{F}$.

Thus existence of a randomization from $\mathcal{E}$ into $\mathcal{F}$ implies that $\mathcal{E}$ is better. The Blackwell–Sherman–Stein theorem says that the converse is also true (under some conditions): if $\mathcal{E}$ is better than $\mathcal{F}$, then there exists a Markov kernel T with $TP_\theta = Q_\theta$ for every $\theta \in \Theta$, that is, $\mathcal{F}$ is a *randomization* of $\mathcal{E}$.

It follows that the comparison of available risk functions and the existence of randomizations lead to the same order on statistical experiments. This is a partial order only, because, in actual fact, there are not so many pairs of experiments that are in an ordered relationship in this way. The success of Le Cam's (1964a) paper was to quantify the comparison idea, so that it becomes an approximation idea. The *deficiency* $\delta(\mathcal{E}, \mathcal{F})$ of the experiment $\mathcal{E}$ relative to the experiment $\mathcal{F}$ is defined as

$$\delta(\mathcal{E}, \mathcal{F}) = \inf_T \sup_{\theta \in \Theta} \|TP_\theta - Q_\theta\|, \tag{7.2}$$

where the infimum is taken over all randomizations T, and the norm $\|\cdot\|$ is the total variation norm (2.4). (The deficiency number or the norm $\|\cdot\|$ are sometimes defined with an additional factor $\frac{1}{2}$ or 2.) A deficiency of zero corresponds to the existence of a perfect randomization, one with the property that $TP_\theta = Q_\theta$ for every $\theta \in \Theta$. Thus, by the Blackwell–Sherman–Stein theorem a deficiency $\delta(\mathcal{E}, \mathcal{F})$ of zero is equivalent to $\mathcal{E}$ being better than $\mathcal{F}$ in terms of risk functions.

However, in most cases the deficiency is strictly positive and hence the Blackwell–Sherman–Stein theorem has nothing to say. Le Cam (1964a) showed that the equivalence between using randomizations and comparison of risk functions extends to the case of two general experiments. He proved the following theorem.

THEOREM 7.1. *For any $\varepsilon \geq 0$ and any loss function with values in the unit interval $\delta(\mathcal{E}, \mathcal{F}) \leq \varepsilon$ if and only if for every procedure τ in $\mathcal{F}$ there exists a procedure σ in $\mathcal{E}$ such that $R(\theta; \sigma) \leq R(\theta; \tau) + \frac{1}{2}\varepsilon$ for every $\theta \in \Theta$.*

This theorem is simple enough and full of statistical meaning: $\frac{1}{2}$ times the deficiency $\delta(\mathcal{E}, \mathcal{F})$ is approximately the maximal difference in risk between $\mathcal{E}$ and $\mathcal{F}$. That Le Cam's paper was held up for a number of years is connected to the fact that the theorem is not quite true if we interpret the objects, experiments, risk, procedures, and deficiency, exactly in the way that we described them so far. It is possible to make somewhat restrictive assumptions on the experiments $\mathcal{E}$ and $\mathcal{F}$

and the decision space $\mathbb{D}$ to save the theorem as it stands, but Le Cam was not willing to do so. In his view it was more elegant and easier to change the meaning of the entities in the theorem. He replaced the Markov kernels, employed both as randomizations and statistical procedures, by certain linear maps, later called "transitions," and appropriately redefined risk functions. In the introduction of his 1964 paper [Le Cam (1964a), page 1419] Le Cam writes:

> The study of definitions of sufficiency is marred by technical difficulties of a measure theoretic nature, which may be judged rather irrelevant for ordinary statistical purposes. To avoid these difficulties we have been led to generalize the usual description of what is meant by an experiment, ignoring σ-additivity and other regularity conditions. The bulk of the paper is intended to show that such a generalization is very convenient in many respects. Furthermore, there is no essential difficulty in returning to the usual system after the main results have been proved.

We shall follow Le Cam on this route in the next section. For now, let us note that the "restrictive assumptions" are actually not very restrictive. Most situations of statistical interest are covered by the following theorem, which is a translation of Le Cam's (1964a) result in more common statistical language. The objects in this theorem may be interpreted in the usual way.

First, assume that the statistical experiments $\mathcal{E} = (\mathcal{X}, \mathcal{A}, P_\theta : \theta \in \Theta)$ and $\mathcal{F} = (\mathcal{Y}, \mathcal{B}, Q_\theta : \theta \in \Theta)$ are dominated, that is, there exist σ-finite measures that dominate all probability measures P_θ and Q_θ, respectively. Second, assume that the sample spaces of $\mathcal{E}$ and $\mathcal{F}$ and the decision space $(\mathbb{D}, \mathcal{D})$ are all Polish spaces with their respective Borel σ-algebras (equivalently: complete separable metric spaces with σ-fields generated by the open or closed balls). Then we define the deficiency $\delta(\mathcal{E}, \mathcal{F})$ by (7.2) with the infimum taken over all Markov kernels T from $(\mathcal{X}, \mathcal{A})$ into $(\mathcal{Y}, \mathcal{B})$. This setting resembles the one of Heyer (1973).

THEOREM 7.2. *Assume the setup of the preceding paragraph. Then* $\delta(\mathcal{E}, \mathcal{F}) \leq \varepsilon$ *if and only if for every Markov kernel* T *from* $(\mathcal{Y}, \mathcal{B})$ *into* $(\mathbb{D}, \mathcal{D})$ *there exists a Markov kernel* S *from* $(\mathcal{X}, \mathcal{A})$ *into* $(\mathbb{D}, \mathcal{D})$ *such that for every loss function* ℓ *with values in* $[0, 1]$, *and every* $\theta \in \Theta$,

$$\int\int \ell(\theta, z) S_x(dz)\, dP_\theta(x) \leq \int\int \ell(\theta, z) T_y(dz)\, dQ_\theta(y) + \tfrac{1}{2}\varepsilon.$$

In a comment on lecture notes distributed at a Yale course in 1994 [organized by David Pollard in honor of Le Cam's 65th birthday; see van der Vaart (1994)] Le Cam wrote that "he had no difficulty with the Polish assumptions, except that they are assumptions." Indeed, it is nice to remove them. A closer look would reveal the different roles the assumptions play and how they could be relaxed. However, a full removal of the assumptions will require that:

- we redefine $\delta(\mathcal{E}, \mathcal{F})$ by taking the infimum over the larger class of "transitions" T

- we replace the Markov kernels S and T by elements of the larger class of "procedures"
- we do not define the risk by an integral or expectation.

Le Cam defended this route in the late 1950s, and never changed his mind on this. For instance, in the first chapter of his 1986 book he points out the arbitrariness of sample spaces in statistics, and eliminates these altogether. Without sample spaces we cannot have Markov kernels, of course.

8. *L* and *M* spaces. What is a statistical experiment according to Le Cam? In his 1964 paper he gives the following definition (page 1421):

> DEFINITION 1. A single stage experiment $\mathcal{E} = \{\Theta, E, \mathcal{X}, \{P_\theta\}\}$ consists of a set Θ, a set E of bounded numerical functions on a set $\mathcal{X}$ and a map $\theta \to P_\theta$ which to each $\theta \in \Theta$ associates a numerical function P_θ defined on E. The system is assumed to satisfy the following requirements.
>
> (i) E is a vector lattice for the usual operations carried out point by point.
> (ii) The function I, identically equal to unity on $\mathcal{X}$, is an element of E.
> (iii) Each P_θ is a positive normalized linear functional on E.
> (iv) E is complete for the norm $\|f\| = \sup\{|f(x)|; x \in \mathcal{X}\}$.

Just before giving this definition Le Cam remarks that the set $\mathcal{X}$ and the nature of E as a space of functions are there "to facilitate a return to the usual structures." Later on he would remove the "sample space" altogether. In his 1986 book an experiment is defined as a map $\theta \mapsto P_\theta$ from an arbitrary set Θ into the nonnegative boundary of the unit ball in an abstract L-space, a Banach lattice or Riesz space of special type.

Le Cam was an early user of the theory of Banach lattices. Some of the essential concepts, including abstract L- and M-spaces, go back to the 1930–1940s, but much of the mathematical theory was developed by functional analysts in the period 1950–1970. The first accounts in book form were those of Schaefer (1974) and Luxemburg and Zaanen (1971).

Whereas a reader of the first chapter of Le Cam (1986) may wonder whether this is a book on statistics at all, the 1964 definition was written purposely in such a way that the statistical experiment is relatively easy to recognize. It is probably a defendable position to say that Le Cam (1964a) did not waste much more effort to appeal to statisticians. He does explain tools and objects step by step as he develops the theory, but he gives few references to the literature, and rarely relates to statistics in a concrete way. [Le Cam's basic reference for Riesz theory probably was the Bourbaki volume on integration, although in his 1964 paper he refers only to the volume of Bourbaki (1955) on topological vector spaces, which appears to have no material on vector lattices.]

The main abstraction is to view a probability measure P_θ not as a prescription for distributing mass on a set $\mathcal{X}$, but rather as a functional that takes functions $f : \mathcal{X} \to \mathbb{R}$ into numbers. If one were given a measurable structure on the set $\mathcal{X}$,

the functions f were chosen measurable, and the object P_θ were a measure relative to this structure, then these functionals could be taken as the maps

$$f \mapsto P_\theta f := \int f(x)\, dP_\theta(x).$$

One may not wish to use all measurable functions here, but may rather single out a collection E of relevant functions, and restrict the P_θ to this collection, as in the definition. As required by (iii) the functionals P_θ are linear and positive in that they take nonnegative functions f into nonnegative numbers. They are thus elements of the *Riesz dual* (or *order dual*) E^* of E, the set of all linear functions $P : E \to \mathbb{R}$ that map *intervals* $[f, g] = \{v \in E : f \le v \le g\}$ into bounded subsets of $\mathbb{R}$.

The order dual is similar to the better known dual space of a normed vector space. In fact, for a partially ordered vector lattice E (as in Le Cam's definition) that is equipped with a norm the order dual and norm dual are identical, provided some mild compatibility properties between norm and ordering exist. However, Le Cam (1964a) did not refer much to the norm structure and worked mostly within the Riesz lattice setup.

In the 1964 paper the preceding definition of an experiment is followed by two more definitions, giving the L- and M-spaces of the experiment $\mathcal{E} = \{\Theta, E, \mathcal{X}, \{P_\theta : \theta \in \Theta\}\}$. The *$L$-space* $L(\mathcal{E})$ is the smallest *band* in E^* that contains $\{P_\theta : \theta \in \Theta\}$, and the *$M$-space* $M(\mathcal{E})$ is the order dual of this band. The concepts of bands and order duals do not make part of standard introductions to functional analysis today and must have been exotic also in the 1950s. We shall not go into the details here, but note that in the case that the P_θ are probability measures on a measurable space, the corresponding L-space is the set of all probability measures that are dominated by some countable linear combination of the P_θ. In particular, the L-space of a dominated experiment $(\mathcal{X}, \mathcal{A}, P_\theta : \theta \in \Theta)$ can be identified with the $L_1(\mathcal{X}, \mathcal{A}, \mu)$-space for μ a dominating measure that is equivalent to the set of P_θ.

It is far from obvious that bands have statistical relevance, but they do. The 1964 paper is truly remarkable, both by translating statistical concepts in concepts of Riesz spaces, and in its level of abstraction and generality. Whereas many mathematical theories evolve in small steps to a higher level, it appears that here Le Cam has made a giant step, which was not announced in earlier work, by himself or others.

Some ten out of the 36 pages are used to "indicate the relations between our description of decision procedures and the more usual ones" [Le Cam (1964a), page 1420]. This number of pages appears high enough, but this number at the same time appears to indicate that the claim that "there is no essential difficulty in returning to the usual system after the main results have been proved" [Le Cam (1964a), page 1419] is perhaps overstated. On the whole the paper was and is difficult to understand. It is doubtful that a paper of this content would receive an enthusiastic reception by the present day board of *The Annals of Statistics*.

It would be senseless to describe Le Cam's framework here in a few pages, but we do wish to include the exact definitions of the deficiency distance, transitions and procedures [see, e.g., Strasser (1985) or Torgersen (1991) for extended introductions]. As we mentioned an experiment $\mathcal{E}$ possesses an "L-space" $L(\mathcal{E})$, which may be taken to be the $L_1(\mathcal{X}, \mathcal{A}, \mu)$-space for a dominating measure if the experiment is represented through probability measures on a measurable space $(\mathcal{X}, \mathcal{A})$ and dominated. Also it does not make much difference for the following to replace the L-space by the set $L(\mathcal{X}, \mathcal{A})$ of all signed measures on $(\mathcal{X}, \mathcal{A})$, which in general will be bigger than $L(\mathcal{E})$. A *transition* from an experiment $\mathcal{E}$ into an experiment $\mathcal{F}$ is a positive, norm-preserving, linear map $T : L(\mathcal{E}) \to L(\mathcal{F})$ between the L-spaces of the experiments. Given representations of the experiments $\mathcal{E}$ and $\mathcal{F}$ through measures on measurable spaces $(\mathcal{X}, \mathcal{A})$ and $(\mathcal{Y}, \mathcal{B})$, a transition can be taken to be such a map between the spaces $L(\mathcal{X}, \mathcal{A})$ and $L(\mathcal{Y}, \mathcal{B})$ of all signed measures on these spaces, or in the dominated cases between the associated L_1-spaces. In this case any Markov kernel T yields a transition $P \mapsto TP$, as in (7.1), but not every transition need be given by a Markov kernel.

The deficiency $\delta(\mathcal{E}, \mathcal{F})$ can now be defined exactly as in (7.2), but with the infimum computed over all transitions T between $\mathcal{E}$ and $\mathcal{F}$. It can be symmetrized to a distance Δ on the collection of experiments by defining

$$\Delta(\mathcal{E}, \mathcal{F}) = \delta(\mathcal{E}, \mathcal{F}) \vee \delta(\mathcal{F}, \mathcal{E}).$$

Le Cam appears never to have given a name to this distance, but often refers to it as the "distance Δ." Several authors now refer to it as the *Le Cam distance*.

One of the nice properties of transitions is that the collection of all transitions possesses certain compactness properties. [To be more precise, the set of all transitions into the second dual of $L(\mathcal{F})$, which is a slightly larger set of operators, is compact for the topology of pointwise convergence.] This implies, for instance, that if the deficiency $\delta(\mathcal{E}, \mathcal{F})$ is zero, then the infimum in (7.2) is attained and hence there exists a transition T such that $TP_\theta = Q_\theta$ for all $\theta \in \Theta$. This is true with Markov kernels only under restrictive assumptions. It can also be shown that the deficiency between two experiments is determined by the deficiencies between all finite subexperiments. In fact

$$\delta(\mathcal{E}, \mathcal{F}) = \sup_I \inf_T \sup_{\theta \in I} \|TP_\theta - Q_\theta\|,$$

where the first supremum is taken over all finite subsets I of Θ. In Section 5 we mentioned that the weak convergence of a sequence of experiments $\mathcal{E}_n$ to a limit $\mathcal{E}$, which can be expressed in convergence of the likelihood ratio processes, is equivalent to the convergence to zero, $\Delta(\mathcal{E}_n^I, \mathcal{E}^I) \to 0$, of the Le Cam distance between all finite subexperiments. In view of the preceding display the difference between weak convergence and convergence in the Le Cam distance ("strong convergence") is uniformity in finite sets.

Just as a Markov kernel has a meaning both as a randomization (i.e., "transition") between experiments and as a decision procedure, a transition plays a double role. In the Wald theory a decision space is a set $\mathbb{D}$, and a statistical procedure is a Markov kernel from the sample space into the decision space. Such a Markov kernel T induces a bilinear map on the product $L(\mathcal{X}, \mathcal{A}) \times F$ of the signed measures on the sample space $(\mathcal{X}, \mathcal{A})$ and a given vector space F of measurable functions $f : \mathbb{D} \to \mathbb{R}$, given by

$$(P, f) \mapsto \int_{\mathcal{X}} \int_{\mathbb{D}} f(y) T(x, dy)\, dP(x).$$

A pair $(P, 1)$ of a probability measure and the unity function is mapped onto 1. Le Cam defines a *procedure* as a bilinear, positive map of norm 1 on $L(\mathcal{E}) \times F$, where F is a uniform lattice of functions on a given set $\mathbb{D}$. We may view such a procedure $(P, f) \mapsto T(P, f)$ also as a map $P \mapsto T(P, \cdot)$ from $L(\mathcal{E})$ into the set of positive, norm-preserving, linear maps on F, in other words as a map from $L(\mathcal{E})$ into the dual space F^* of F. Because this dual space is necessarily an L-space, we can also think of T as a transition of $L(\mathcal{E})$ into F^*.

Because we have now lost both the law of the observation and the induced sample distribution of the Markov kernel, we cannot define risk as an expectation. However, the map $(P, f) \mapsto T(P, f)$ is the abstraction of the integral in the preceding display. If $f = \ell_\theta = \ell(\theta, \cdot)$ were our loss function, then the map $\theta \mapsto T(P_\theta, \ell_\theta)$ would be the natural candidate risk function of the procedure T. It may well be that the loss functions ℓ_θ are not in the domain F of the procedure. To accommodate this possibility, the risk function of the procedure T is defined as

$$R(\theta; T) = \sup_{0 \le f \le \ell_\theta,\, f \in F} T(P_\theta, f).$$

This can be defined for an arbitrary loss function ℓ: no measurability or integrability is required. On the other hand, if we are in a setting of measures on a sample space and procedures are given by Markov kernels, then the new definition of "risk" will reduce to the old definition of risk as an expectation only under some regularity conditions (e.g., lower semicontinuity if F is taken to be the set of bounded, uniformly continuous functions on a metric space).

It is now possible to reread Theorem 7.1. It is correct if the deficiency, risks and procedures are interpreted according to the preceding definitions.

In general, the set of procedures is larger than the set of Markov kernels. A simple example illustrating this is the statistical problem of testing a null hypothesis that an observation X was sampled from the uniform distribution on $[0, 1]$ versus the alternative that it was sampled from one of the Dirac measures δ_θ, for $\theta \in [0, 1]$. We can take the two-point set $\{0, 1\}$ as a decision space. A Markov kernel from $[0, 1]$ into $\{0, 1\}$ corresponds to the test that rejects the null hypothesis with probability $T_x\{1\}$ if x is observed. The error probabilities of this test are

$$\alpha = \int_0^1 T_x\{1\}\, dx, \qquad \beta = \sup_{0 \le \theta \le 1} \int (1 - T_x\{1\})\, d\delta_\theta(x) = \sup_{0 \le \theta \le 1} (1 - T_\theta\{1\}).$$

Of course, it is not possible to test the two hypotheses in a sensible way: every test has $\alpha + \beta \geq 1$. However, using the Hahn–Banach theorem one can show that there exists a transition T from the set of all signed Borel measures on $[0, 1]$ to the dual $F^* = \mathbb{R}^2$ of $F = C(\{0, 1\}) = \mathbb{R}^2$ such that

$$T\big(U[0, 1], (0, 1)\big) = 0, \quad 1 - T\big(\delta_\theta, (0, 1)\big) = 0 \qquad \text{every } \theta \in [0, 1].$$

These numbers give the risk function of the transition and hence must be interpreted as its error probabilities. It follows that there exists a perfect transition, even though there is no sensible statistical procedure. [The notation $(0, 1)$ is used to denote the function $f : \{0, 1\} \to \mathbb{R}$ with $f(0) = 0$ and $f(1) = 1$. This is the loss function under the null hypothesis, the uniform measure. The loss function under the alternative hypothesis, the Dirac measures, is $(1, 0)$. We must have $T(P, (1, 0)) = 1 - T(P, (0, 1))$, because $T(P, (1, 1)) = 1$ for every transition.]

This is bothersome. The discrepancy between the abstract setting and the measure-theoretic formulation appears to arise, because a Markov kernel is more than just a positive, linear map. Apparently, more than intuition may be lost when viewing a probability measure as an algebraic object, rather than as a recipe to distribute mass. In the preceding example the discrepancy arises because the experiment is not dominated. The example appears to indicate that it is not enough to solve a statistical problem within the abstract formulation, but it is necessary to translate the solution back into statistical terms. Le Cam had no difficulty in switching from the abstract to the measure-theoretic setup and vice versa. The fact that he rarely makes the transition explicit in his writings is one reason that his work is hard to read.

Is the preceding example a serious challenge to Le Cam's theory? It is not a serious statistical example, because it poses a problem that has no solution in the usual statistical framework. Thus one could argue that Le Cam's abstract setting simply allows the discussion and "solution" of a problem that is outside the scope of the usual framework. This position was taken by David Pollard (personal communication), who also pointed out that allowing transitions as statistical procedures is similar to accepting Kolmogorov's definition of a conditional expectation in cases where there is no regular version of a conditional law, a widely accepted practice. On the other hand, the example does appear to warn us not to stop after solving a problem in Le Cam's framework, but to pursue its implications for the usual setup.

While Le Cam would acknowledge a role for measure theory, his main objection to the usual way of describing statistical experiments is that a given practical situation might be describable by many different types of sample spaces and "true" measures. If one happened to choose the "wrong one," one might get burdened by technical problems, for no good reason. Furthermore, any experiment in Le Cam's sense can be represented as an experiment in the usual way if the sample space is chosen appropriately [Le Cam (1986), pages 12–13]:

> Thus, as asserted previously, we have not introduced in the abstract framework any objects which could not be introduced as well in the traditional structure.
>
> Thus one could ask "why then use the abstract framework?" The point is that the above representation is very special. It is only one of a multitude of possibilities, and the usual setup where $\mathcal{E} = \{P_\theta, \theta \in \Theta\}$ is given by probability measures P_θ on a σ-field $\mathcal{A}$ carried by a set $\mathcal{X}$ does not insure that the σ-field $\mathcal{A}$ or the set $\mathcal{X}$ are selected well enough to be able to proceed without trouble. The abstract framework avoids the troubles caused by $\mathcal{X}$ or similar sets by ignoring them. However, since many readers may be much more familiar with the traditional setup, we shall now indicate that it may be used *provided* suitable assumptions are duly satisfied.

After discussing such suitable assumptions, Le Cam [(1986), pages 14–15] continues:

> Returning to the why and how of the "abstract" setting given previously, the reader will easily convince himself that the abstract setting is just meant to ignore the representation problems described here. The main "statistical" statements elaborated in this book do not depend on such representations.

One may or may not go along with this philosophical statement. A fact is that for many mathematical proofs it certainly works to solve the problem at a higher level of abstraction first, and next try to translate it back. This was a major point of Le Cam (1964a), repeated in most of his later work. In this review we have spent relatively little time on the abstract setup, but instead have focused on a selection of other great ideas, more closely connected to statistics as a part of probability. It is likely, however, that Le Cam discovered these great ideas from his abstract viewpoint first. David Pollard (personal communication) pointed out that for Le Cam, trained in the Bourbaki tradition, the "abstract point of view" might have been perfectly obvious and natural, so that representations in terms of measures always were a secondary and perhaps even unnecessary step.

In any case, one must agree that a result such as Theorem 7.1 is mathematically appealing.

9. Comparison of experiments. Even though the deficiency distance has strong statistical significance, it has turned out to be hard to use it, because it is difficult to compute for a given pair of experiments. Among the rare situations where a more or less concrete representation is possible are group models. For instance, the deficiency between two full shift experiments $\mathcal{E} = (P * \delta_h : h \in \mathbb{R}^k)$ and $\mathcal{F} = (Q * \delta_h : h \in \mathbb{R}^k)$ for Lebesgue absolutely continuous probability measures P and Q on $\mathbb{R}^k$ can be written as

$$\delta(\mathcal{E}, \mathcal{F}) = \inf_M \|P * M - Q\|, \tag{9.1}$$

where the infimum is taken over all probability measures M. The explanation for this is given already in Le Cam (1964a): the transitions between a pair of experiments that are invariant under the action of a group can be chosen to be

invariant under this action, without changing the value of the infimum defining the deficiency. In particular, for shift experiments the transitions can be chosen invariant under translation. It can next be shown that all translation invariant transitions are representable through convolution in the form $TP = P * M$, for some measure M.

Le Cam (1964a) proved this invariance property for group models using the Kakutani fixed point theorem, and later applied the same arguments in his treatment of Hájek's convolution theorem and its generalizations, including the infinite-dimensional convolution theorem of Le Cam (1994). His approach also led to rigorous, general statements of the Hunt–Stein theorem, and other results that are well known, but rarely rigorously stated.

If P and Q in the preceding are Gaussian measures, then it can be shown that the minimizing M in (9.1) can also be chosen Gaussian and hence it is possible to compute the deficiency exactly. More generally, Torgersen studied deficiencies between experiments corresponding to (Gaussian) linear models. [See Torgersen (1991).] These are among the rare examples where the deficiency is known exactly.

Unlike Torgersen, whose book is titled "Comparison of Statistical Experiments," Le Cam (1986) presented his book with asymptotics in mind: "Asymptotic Methods in Statistical Decision Theory." The scope for the deficiency distance is considerably wider for asymptotic comparison of experiments. The purpose may then be to show that $\Delta(\mathcal{E}_n, \mathcal{F}_n) \to 0$ for two seemingly very different sequences of experiments. The two sequences $\mathcal{E}_n$ and $\mathcal{F}_n$ are then called "equivalent." Risk functions available in $\mathcal{E}_n$ must then also be approximately available in $\mathcal{F}_n$ as $n \to \infty$.

As noted in Section 5, for experiments with finitely many parameters the weak convergence $\mathcal{E}_n \to \mathcal{E}$ is equivalent to (or defined by) $\Delta(\mathcal{E}_n, \mathcal{E}) \to 0$. This immediately yields many examples of equivalent sequences of experiments.

Another easy example of asymptotically equivalent sequences of experiments arises from local asymptotic normality. For a LAN sequence $(P_{n,\theta} : \theta \in \Theta)$ and every compact $K \subset \mathbb{R}^k$,

$$\Delta\big((P_{n,\theta+\delta_n h} : h \in K), \big(N(h, J_\theta^{-1}) : h \in K\big)\big) \to 0. \tag{9.2}$$

This goes only slightly beyond the weak convergence of the sequence of experiments $(P_{n,\theta+\delta_n h} : h \in \mathbb{R}^k)$ to its Gaussian limit, which is equivalent to convergence of all finite subexperiments in the Le Cam distance. In (9.2) the finite sets of parameters are replaced by compact sets.

The asymptotic equivalence (9.2) suffers from the same drawback as the weak convergence of experiments connected to LAN, in that it gives a local equivalence, for the original experiments rescaled around a fixed parameter value θ. Generally the result is valid for every θ and one may next try to glue the different approximations together into a global one. This was achieved by Le Cam (1975a, 1985a) for independent observations under the assumption of existence of a $\sqrt{n}$-consistent sequence of estimators. The two-step approach, local approximation

followed by gluing together using initial estimators, is reminiscent to the one in his 1960 paper on local asymptotic normality, but it differs at essential points. In fact, the 1960 paper predates the introduction of the deficiency distance and hence could not make reference to comparison of experiments. Le Cam [(1985a), page 234] restates the LAN conditions, roughly requiring:

(A0) existence of $\sqrt{n}$-consistent estimators for θ;
(A1) the approximation (9.2) holds at every θ.

In fact, he imposes conditions (A0), (A1), (A2) more general than this, allowing the parameter set to be a general metric space, not necessarily a subset of Euclidean space. The additional condition (A2) then roughly requires that the local parameter set is approximable by a finite-dimensional set. He remarks [Le Cam (1985a), page 234)]:

> Les conditions LAN de [Le Cam 1960a] ne sont pas énoncées sous cette forme pour une raison très simple: Elles ont été écrites en décembre 1957 alors que la distance Δ n'a été introduite qu'en décembre 1958. Toutefois on peut s'assurer sans grande difficulté que la construction d'estimateurs asymptotiquement exhaustifs de [Le Cam (1960a)] dépend essentiellement seulement des conditions (A0), (A1) et (A2). Elle dépend naturellement beaucoup de (A0). Cette condition a disparu dans ce qu'il est maintenant convenu d'appeler les conditions LAN (voir par exemple Ibragimov–Has'minskii). Sans une condition telle que (A0) on ne peut espérer obtenir que des résultats tout à fait locaux.

For our simplistic LAN situation (2.1), with an open parameter set $\Theta \subset \mathbb{R}^k$, one type of global approximation can be obtained as follows. Because the normal distribution depends continuously on its covariance matrix relative to the total variation distance, the normal distributions $N(h, J_\theta^{-1})$ in (9.2) can be replaced by $N(h, J_{\theta+\delta_n h}^{-1})$ if the matrices J_θ^{-1} depend continuously on the parameter θ. It is a trivial consequence of the definition of the Le Cam distance and the continuity of the normal distributions in its parameters, that the asymptotic equivalence (9.2) then remains valid. (As transitions for comparison of the two normal experiments we can use the identity, in both directions.) We obtain an approximation by the heteroscedastic Gaussian experiments $(N(h, J_{\theta+\delta_n h}^{-1}) : h \in K)$. This more complicated approximation has the advantage that the special role of the central parameter θ diminishes if we rescale back to the original parameter: we may write the approximation as

$$\Delta\big((P_{n,\theta'} : \theta' \in \theta + \delta_n K), \big(N(\theta', \delta_n^2 J_{\theta'}^{-1}) : \theta' \in \theta + \delta_n K\big)\big) \to 0. \tag{9.3}$$

[The Gaussian experiment in this display is equivalent to the Gaussian experiment $(N(\delta_n^{-1}(\theta' - \theta), J_{\theta'}^{-1}) : \theta' \in \theta + \delta_n K)$, by sufficiency.] This shows that the sequences of experiments $(P_{n,\theta} : \theta \in \Theta)$ and $(N(\theta, \delta_n^2 J_\theta^{-1}) : \theta \in \Theta)$ are asymptotically equivalent if restricted to (small, shrinking) subsets of the parameter set Θ. In general, this does not imply anything about the equivalence of the full experiments. However, Le Cam (1975a) proved that these local equivalences and in addition the

existence of estimators that can reduce the global problem to the correct local problem imply the global equivalence. In the present case we need estimators $\hat{\theta}_n$ that are δ_n^{-1}-consistent uniformly in θ.

We shall also need that the LAN approximations (9.2) are valid uniformly in θ, for every compact set K, which we shall refer to as *uniform LAN*. [Such uniformity follows from uniform versions of the LAN expansion (2.1).]

THEOREM 9.1. *Suppose that the sequence of experiments* $(P_{n,\theta} : \theta \in \Theta \subset \mathbb{R}^k)$ *is uniformly LAN with invertible matrices* J_θ^{-1} *that are norm-bounded and depend uniformly continuously on the parameter. Assume that there exist estimators* $\hat{\theta}_n$ *such that* $\mathrm{P}_\theta(\|\hat{\theta}_n - \theta\| > M_n\delta_n) \to 0$ *uniformly in* θ, *as* $n \to \infty$, *for every sequence* $M_n \to \infty$. *Then the sequences of experiments* $(P_{n,\theta} : \theta \in \Theta)$ *and* $(N(\theta, \delta_n^2 J_\theta^{-1}) : \theta \in \Theta)$ *are asymptotically equivalent.*

Surprisingly, a theorem of this type cannot be found in Le Cam (1975a). Most of this paper is concerned with proving the existence of estimators $\hat{\theta}_n$ satisfying the condition of the theorem, in the case of independent observations. One of the seven sections addresses the global equivalence. It contains the key theorem that glues the local approximations together given the estimator sequence $\hat{\theta}_n$, but this theorem is not applied to any concrete case. This is odd, because in the introduction to the paper Le Cam motivates the search for conditions for existence of suitable preliminary estimators $\hat{\theta}_n$ exactly by the fact that they are needed for global approximations. Le Cam [(1975a), page 13]:

> The idea that "when the number of observations is large" such a family can be approximated by a Gaussian family of distributions occurs in the remarkable paper Wald (1943). Subsequently the present author suggested that asymptotically sufficient estimates providing the kind of approximation described by Wald can often be obtained by a two steps "adaptive" procedure as follows. One first finds a good but rough estimate $\hat{\theta}$ of the value of θ. Then, in the vicinity of the estimated value, one approximates the logarithms of likelihood ratios by a suitable quadratic expression. One proceeds as if the quadratic expression was obtained from the logarithms of likelihood ratios of an actual Gaussian family of measures.

Perhaps Le Cam found the preceding theorem too obvious a corollary to state it, or perhaps he did not want to state a corollary that was not in its final form. He must have known the theorem already in 1975.

Le Cam took up the problem of global approximation again in Le Cam (1985a). A good part of this paper is concerned with generalizations of (9.2) to situations where the parameter set Θ is allowed to be a general metric space, subject to certain dimensionality restrictions, and hence where there is no nice finite-dimensional, linear, local parameter set. Some of these generalizations concern models that are infinite-dimensional in the usual sense. In other examples Θ is Euclidean, but the statistical model requires an infinite-dimensional Gaussian approximation.

One particular example of the second type is the location model generated by the density $p(x) = c\exp(-|x|^\alpha)$ for $0 < \alpha < 1/2$. If $P_{n,\theta}$ is the distribution of an i.i.d. sample of size n from the density $x \mapsto p(x-\theta)$, then the local experiments $(P_{n,\theta+\delta_n h} : h \in K)$, for $\delta_n = n^{-1/(1+2\alpha)}$ and a compact interval $K \subset \mathbb{R}$, converge (in the Le Cam distance) to the experiment consisting of observing a Gaussian process $(X_t : t \in K)$ with continuous sample paths and

$$\mathrm{E}_h X_t = C_\theta\big(|h|^{1+2\alpha} - |t|^{1+2\alpha} - |h-t|^{2\alpha+1}\big),$$
$$\operatorname*{cov}_h(X_s, X_t) = C_\theta\big(|s|^{1+2\alpha} - |t|^{1+2\alpha} - |s-t|^{2\alpha+1}\big).$$

An intuitive explanation is that the process X is the limit in distribution of the log likelihood ratio process under $P_{n,\theta+\delta_n h}$ [see Prakasa Rao (1968), Pflug (1983) or Janssen, Milbrodt and Strasser (1985)]. The convergence of experiments is noted in Le Cam [(1969), pages 110–111]. The Gaussian process X is highly nonlinear in the arguments (h,t). The resulting Gaussian experiment is a (nonlinear) one-dimensional curve within an infinite-dimensional Gaussian experiment.

We shall turn to Le Cam's 1985 result on global approximation. Let $(P_\theta : \theta \in \Theta)$ be an experiment with arbitrary parameter set Θ. For simplicity we assume that it is dominated by a σ-finite measure μ. Let H be the Hellinger distance, defined by its square

$$H^2(\theta,\theta') = \tfrac{1}{2}\int \big(\sqrt{p_\theta} - \sqrt{p_{\theta'}}\big)^2\, d\mu. \tag{9.4}$$

Assume that, for any sequence $(\theta_n, \theta_n') \in \Theta^2$ such that $nH^2(\theta_n,\theta_n')$ is bounded and every $\varepsilon > 0$,

$$nP_{\theta_n}\left|\sqrt{\frac{dP_{\theta_n'}}{dP_{\theta_n}}} - 1\right|^2 1\left\{\left|\sqrt{\frac{dP_{\theta_n'}}{dP_{\theta_n}}} - 1\right| > \varepsilon\sqrt{n}\right\} \to 0,$$
$$nP_{\theta_n'}(dP_{\theta_n} = 0) \to 0. \tag{9.5}$$

These conditions replace the LAN condition. The first condition in the display is a Lindeberg condition that ensures that the variables $\sqrt{dP_{\theta_n'}/dP_{\theta_n}} - 1$ satisfy the central limit theorem. In view of Le Cam's second lemma (see Section 12) this gives the asymptotic normality of the sequence of log likelihood ratios.

We define the dimension numbers "$D(\tau)$ of Θ for the pair (H,τ)" in Section 10. For an initial understanding of the following theorem it suffices to know that they are uniformly bounded if Θ under H is metrically smaller than a subset of a Euclidean space, for instance if $k\|\theta-\theta'\|^\alpha \le H(\theta,\theta') \le K\|\theta-\theta'\|^\alpha$ for some positive constants k, K and $\alpha > 0$.

THEOREM 9.2. *Assume that* (9.5) *holds and that the dimension* $D(\tau)$ *of* Θ *for the pair* (H,τ) *is uniformly bounded in* τ. *Then there exists a sequence of Gaussian experiments* $\mathcal{F}_n$ *with parameter set* Θ *that is asymptotically equivalent to the sequence* $\mathcal{E}_n = (P_\theta^n : \theta \in \Theta)$.

Le Cam [(1985a), Théorème 4.3] proves a somewhat more general result, allowing the parameter set Θ to depend on n, and experiments with an arbitrary number of observations, which need not be identically distributed.

In general, the Gaussian experiments $\mathcal{F}_n$ do not have a simple description, as in Theorem 9.1. Le Cam defines an experiment $(Q_\theta : \theta \in \Theta)$ to be *Gaussian* if the measures Q_θ are mutually absolutely continuous and the log likelihood ratios process $\log(dQ_\theta / dQ_{\theta_0})$ is a Gaussian process under Q_{θ_0} (for some θ_0 and then for all θ_0). Every Hilbert space V indexes a "canonical" Gaussian experiment $(G_v : v \in V)$, and it can be shown that any Gaussian experiment arises as a subexperiment of such a canonical experiment. Taking this for granted for the moment we can describe the experiments $\mathcal{F}_n$.

Le Cam (1985a) constructs the experiments $\mathcal{F}_n$ in the preceding theorem as follows. Let V be the set of all discrete signed measures with finitely many support points on Θ. We can define inner products on V by

$$\langle v_1, v_2 \rangle_n = 4n \iiint \sqrt{p_\theta} \sqrt{p_\eta}\, d\mu\, dv_1(\theta)\, dv_2(\eta).$$

For the semi-pre-Hilbert space V we can construct a canonical experiment $(G_{n,v} : v \in V)$ (see below). Next we embed Θ into V through the map $\theta \mapsto \delta_\theta - \delta_{\theta_0}$, for some arbitrary $\theta_0 \in \Theta$, where δ_θ is the Dirac measure at θ. This defines $\mathcal{F}_n$ as the subexperiment $\mathcal{F}_n = (G_{n,\delta_\theta - \delta_{\theta_0}} : \theta \in \Theta)$.

In order to define the canonical experiment $(G_{n,v} : v \in V)$ let $(Z_{n,v} : v \in V)$ be an isonormal Gaussian process indexed by $(V, \langle \cdot, \cdot \rangle_n)$: a mean zero Gaussian process with covariance $\operatorname{cov}(Z_{n,v_1}, Z_{n,v_2}) = \langle v_1, v_2 \rangle_n$. If this Gaussian process is defined on the probability space $(\Omega, \mathcal{U}, G_{n,0})$, then the canonical Gaussian experiment can be defined as $(\Omega, \mathcal{U}, G_{n,v} : v \in V)$ with $G_{n,v}$ defined through its density relative to $G_{n,0}$ given by

$$\frac{dG_{n,v}}{dG_{n,0}} = e^{Z_{n,v} - (1/2)\|v\|_n^2}.$$

This is very general, but also somewhat abstract. In particular, we do not immediately regain the simple Gaussian experiments with Euclidean sample spaces, which arise from LAN. This results from the fact that a given experiment may have many representations: two experiments $\mathcal{E}$ and $\mathcal{F}$ that do not look alike at all may well be equivalent in the sense that $\Delta(\mathcal{E}, \mathcal{F}) = 0$.

Another way of describing the canonical Gaussian experiment would be to say that it consists of observing the process $(X_{n,w} : w \in V)$ for $X_{n,w} = Z_{n,w} + \langle w, v \rangle_n$, where v is the unknown parameter. The marginal distributions of this process are multivariate normal with covariance matrix the identity and the parameter v appearing only in the mean vector, in a linear fashion. Thus the experiment can be termed linear, homoscedastic Gaussian. Any Gaussian experiment can be written in this form, for some Hilbert space V and the parameter ranging over some subset of V.

In that particular representation the observation takes its values in the sample space $\mathbb{R}^V$, which is of very high dimension. To reduce the dimensionality of the observation we may restrict the index to a subset W of V. In particular, we might use a subset $W \subset V$ that permits an isonormal Gaussian process $(Z_{n,w} : w \in W)$ with bounded, uniformly continuous sample paths, relative to the norm on $W \subset V$. If the closed linear span of W contains V, then we can also describe the canonical experiment as consisting of observing $(Z_{n,w} + \langle w, v\rangle_n : w \in W)$. This process takes its values in the sample space $UC(W)$ of uniformly continuous functions on W, and can be termed a Brownian motion process with linear drift. If we are interested in the restriction of the experiment to a subset $V_0 \subset V$ of parameters, we can reduce W accordingly to a subset whose closed linear span contains V_0, down to a finite set if V_0 is finite dimensional.

Because $\langle v_1, v_2\rangle_n = n\langle v_1, v_2\rangle_1$ for every $v_1, v_2 \in V$, the process $Z_{n,v}$ can be constructed as $\sqrt{n}Z_{1,v}$. By rescaling the observation we can also describe the Gaussian experiment as observation of the process $(n^{-1/2} + Z_{1,w} + \langle w, v\rangle_1 : w \in W)$.

Thus the approximating Gaussian experiments can be described in many ways. Which representation is the most useful one depends on the purpose of the approximation. Linear, homoscedastic representations may seem easiest to handle, but if the dimensionality of the Gaussian experiment is large, or the parameter is restricted to a nonlinear subset, classical decision theory for Gaussian experiments may be insufficient to yield the desired results. It is fortunate that LAN leads to such simple Gaussian experiments, at least locally.

The experiments $(N(\theta, \delta_n^2 J_\theta^{-1}) : \theta \in \Theta)$ are, in general, not Gaussian according to Le Cam's definition. They are (curved) exponential families of degree two. In Chapter 14 of his 1986 book, Le Cam studies approximations by exponential families, without, according to his own words, reaching the same depth as for the asymptotically Gaussian case.

Even though the preceding theorem also covers certain infinite-dimensional situations, Le Cam's (1975a, 1985a) main focus appears to be the finite-dimensional models. It was not until recently that it was discovered that similar techniques can be applied to the usual infinite-dimensional statistical models of interest. A breakthrough was obtained by Nussbaum (1996) following work by Brown and Low (1996), and this has motivated several developments in the past years. We do not intend to review these developments, but we do include a discussion of Nussbaum's result, as it is an important recent motivation for studying Le Cam's writings.

Nussbaum considers the problem of estimating a density f on the unit interval $[0, 1]$ based on an i.i.d. sample of size n. The density is known to be bounded above and below by given constants and to satisfy a uniform Lipschitz condition of order $\alpha > 1/2$: for some L,

$$|f(x) - f(y)| \le L|x - y|^\alpha. \tag{9.6}$$

The sequence of experiments $\mathcal{E}_n$ obtained by letting $n \to \infty$ can be approximated by a sequence of Gaussian experiments. As we noted, Gaussian experiments can be represented in many equivalent ways. In the present case a convenient concrete representation is as follows. Let the experiment $\mathcal{F}_n$ consists of observing a stochastic process $(X_t : t \in [0, 1])$ taking its values in the space $C[0, 1]$ and possessing the same law as

$$\int_{[0,t]} f^{1/2}(s)\, ds + \tfrac{1}{2} n^{-1/2} W_t,$$

for W_t a standard Brownian motion process.

THEOREM 9.3. *Let the experiments $\mathcal{E}_n$ and $\mathcal{F}_n$ be as described previously with parameter set equal to the collection of all probability densities f on the unit interval $[0, 1] \subset \mathbb{R}$ satisfying the Lipschitz condition* (9.6) *and* $\min_x f(x) \geq \eta$ *for given positive constants η, L. Then $\Delta(\mathcal{E}_n, \mathcal{F}_n) \to 0$ as $n \to \infty$.*

Nussbaum's proof of this result follows the same principle as Le Cam's proof of the preceding theorems. First it is shown that for every fixed f a local experiment around f can be approximated by a Gaussian experiment. Next the local approximations are glued together in a global approximation through the use of an estimator with an appropriate convergence rate.

A closer look reveals two important differences. First the local experiments used in Theorems 9.1 and 9.2 are within the range of contiguity, which means not bigger than of the order $1/\sqrt{n}$ in the Hellinger distance, in the case of i.i.d. observations. This is good enough in that case, because there exist initial estimators that attain this rate of convergence, for instance in view of Theorem 10.2 discussed in the next section. In truly infinite-dimensional situations the best estimators do not possess such a fast rate of convergence. As a result it is necessary to obtain equivalence for larger local experiments, before proceeding to the second part of the argument, pasting the local experiments together. A second augmentation of Le Cam's approach is in the pasting argument. Le Cam's argument [e.g., Proposition 8, page 78 of Le Cam (1986), or Theorem 1 on page 23 of Le Cam (1975a)] is completely general, without requiring an i.i.d. setup, but does appear to work only for models of bounded metric dimension. Nussbaum's argument uses the i.i.d. nature of the experiment by splitting the experiment in two independent halves and works without dimensionality restrictions.

The importance of Theorem 9.3 lies, besides in its intellectual interest, in that it relates seemingly different experiments in a very strong sense. The Gaussian experiments do involve observation of a drifted Brownian motion and may not be as simple as one might wish. However, the parameter n, which has the complicated role of product size in the original experiments, enters the Gaussian experiments in a transparent way as a scalar measuring the noise level. Moreover, computations for the Gaussian model, which can be reduced to independent

Gaussian variables by sufficiency, are relatively easy. In any case, for several models for the parameter f of interest (e.g., Hölder or Besov balls) the "Gaussian white noise model" had been studied prior to 1996 for many years, particularly by the Russian school in statistics [e.g., Ibragimov and Khasminskii (1977), Pinsker (1980)]. The theorem allows one to transfer known results for the Gaussian experiments, such as minimax bounds and constants, to the problem of density estimation, under some conditions. More recent work has achieved the same for other problems besides density estimation.

10. Metric entropy. In the last decade a number of authors have worked on nonparametric or semiparametric estimation from the perspective of complexity of a statistical model. These authors were interested in minimum contrast estimators [e.g., van de Geer (1993), Wong and Shen (1995), Birgé and Massart (1993, 1998)] or posterior distributions [e.g., Ghosal, Ghosh and van der Vaart (2000)] and complexity is measured through entropy numbers. The use of entropy numbers to bound the risk of estimators goes back to Le Cam, who introduced metric dimension in statistics in Le Cam (1973a), made significant progress in Le Cam (1975a), and was a great inspiration for further work in France, in particular that of Birgé (1983).

Le Cam himself did not have great interest in minimum contrast estimators (compare the remarks on his attitude to maximum likelihood in Section 11), but focused on obtaining upper and lower bounds for rates of convergence of estimators or posterior distributions, expressed in the metric entropy of the parameter set. The title of his 1973 paper "Convergence of estimates under dimensionality restrictions" reflects his view of metric entropy as a means of moving away from the usual parametric models to more general models, still restricted by dimensionality. Birgé (1983) showed that the approach could lead to completely general results that can cover both parametric models, where the dimension is fixed and the rate of convergence for a canonical distance is "always" the square root of the number of observations, and general nonparametric statistical models, such as classes of densities restricted by smoothness.

A concept of metric entropy was defined by Kolmogorov and Tikhomirov [(1961); Russian original (1959)], and was studied subsequently for numerous metric spaces. This concept was used by Dudley (1967) to give sufficient conditions for continuity of Gaussian processes, and was the basis for striking generalizations of Donsker's theorem on the weak convergence of the empirical process. For the statistical purpose mentioned previously the concept can be used in its original form, but it is not quite the right concept. Instead Le Cam introduced a concept of *metric dimension*, which could be described as a "local entropy."

DEFINITION 10.1. Given a positive number τ the *dimension* $D(\tau)$ of a metric space Θ with metric H *for the pair* (H, τ) is the smallest number D such that for every $\delta \geq \tau$ every set of diameter 2δ can be covered by no more than 2^D sets of diameter δ.

According to this definition we may obtain many "dimensions" $D(\tau)$ of Θ, if we vary the number τ. To understand this, it is useful to consider the case of a subset Θ in $\mathbb{R}^k$, where, for simplicity we take the supremum norm $\|\theta\|_\infty = \max_i |\theta_i|$. A k-dimensional ball (i.e., hypercube) of radius δ can be covered by 2^k hypercubes of radii $\delta/2$. (We just cut each of the k axes in two.) It follows that, for any $\tau > 0$, the dimension $D(\tau)$ of a set $\Theta \subset \mathbb{R}^k$ for the pair $(\|\cdot\|_\infty, \tau)$ is bounded above by k. Thus Le Cam's dimension numbers are bounded above by the "true" dimension of the set in this case.

The title of Section 16.5 in Le Cam (1986), from where we copied the preceding definition and the following results, is "Estimates for Finite Dimensional Parameter Spaces." This may be somewhat misleading, because the results concern the situation where the dimension numbers $D(\tau)$ are finite for every τ, but not necessarily bounded. Finiteness of the function $\tau \mapsto D(\tau)$ implies some restrictions on the size of Θ, but allows Θ to be infinite-dimensional in the usual sense.

This becomes clearer if the dimension numbers are related to Kolmogorov's metric entropy. Let $N(\delta, \Theta, H)$ be the minimal number of balls of radius $\delta > 0$ needed to cover the set Θ. Every set of diameter 2δ fits into a ball of radius 2δ, and a covering of such a ball by balls of radii $\delta/2$ gives a covering by sets of diameter δ. Thus

$$D(\tau) \le \sup_{\theta\in\Theta} \sup_{\delta\ge\tau} N\bigl(\delta/2, \{\theta' \in \Theta : H(\theta', \theta) \le 2\delta\}, H\bigr) \le N(\tau/2, \Theta, H).$$

The right-hand side is finite for every $\tau > 0$ if and only if the set Θ can be covered by finitely many balls of an arbitrarily small radius. This is equivalent to the compactness of the metric completion of Θ. Many sets Θ that are ordinarily considered to be infinite-dimensional meet this criterion. For example, for the set of all monotone densities $\theta : [0, 1] \to [0, 10]$ and H the Hellinger distance the right side of the display is of the order $1/\tau$ as $\tau \to 0$, whereas for densities $\theta : [0, 1] \to \mathbb{R}$ whose root has α uniformly bounded derivatives it is of the order $(1/\tau)^{1/\alpha}$. Finiteness of Le Cam's dimension numbers requires less than compactness.

We state one typical theorem showing the existence of certain estimators attaining a rate of convergence that is upper bounded through Le Cam's dimension numbers. For clarity we state the theorem for the case of i.i.d. observations, so that it gives a rate in terms of the number of observations, and will be concerned with a rate of convergence only. The theorem can be found in Le Cam [(1986), page 498 or 505] in a more general situation and as an explicit upper bound on the risk.

We look for an estimator $\hat{\theta}_n$ for a parameter θ contained in an arbitrary set Θ based on n i.i.d. observations from a density p_θ relative to a given measure. The metric of choice is the *Hellinger distance*, whose square is given in (9.4). This particular distance is chosen, because it guarantees the existence of certain tests. Other distances could be used, but the tests should be derived explicitly and the result will be particular to the sitation at hand [cf. Donoho and Liu (1991) for examples].

THEOREM 10.2. *For every sequence of numbers $\varepsilon_n \downarrow 0$ such that $D(\varepsilon_n) \leq n\varepsilon_n^2$ for every n there exist estimators $\hat{\theta}_n$ and a constant C such that $\mathrm{E}_\theta H^2(\hat{\theta}_n, \theta) \leq C\varepsilon_n^2$ for every $\theta \in \Theta$.*

For instance, for truly finite-dimensional models of dimension k we derive the rate $\varepsilon_n = k/\sqrt{n}$, because $D(\varepsilon_n) \leq k = n(k/\sqrt{n})^2 = n\varepsilon_n^2$. For a big model with local entropy of the order $(1/\varepsilon)^{1/\alpha}$ we obtain the rate $\varepsilon_n = n^{-\alpha/(2\alpha+1)}$, because this is the minimal solution of the equation $(1/\varepsilon_n)^{1/\alpha} \leq n\varepsilon_n^2$. These are rates for the Hellinger distance. For irregular parametric models these may well translate into different rates (also faster than $1/\sqrt{n}$) for the parameter in a natural distance on the parameter set. Characteristically, concrete examples of models and the corresponding rates are not given by Le Cam (1986) or Le Cam (1975a). It is necessary to read Birgé (1983, 1986) to be able to appreciate what is being done here. According to Birgé (personal communication), Le Cam was aware of the applications to infinite-dimensional problems as early as the 1970s, but this is not readily apparent from Le Cam's writing.

Estimators as in the preceding theorem are constructed explicitly by Le Cam and Birgé (1983), but appear not to be very practical. The recent work mentioned at the beginning of this section shows that similar results can be obtained for more natural estimators defined through a minimization criterion, such as least squares or maximum likelihood. The drawback of such "natural" approaches is that they only can be proved to work under more restrictive conditions. For instance, the simple condition on the metric dimension of the theorem is replaced by an integral condition involving the larger bracketing entropy of the form

$$\int_0^{\varepsilon_n} \sqrt{\log N_{[]}(\varepsilon, \{\theta : H(\theta, \theta_0) \leq 2\varepsilon\}, H)}\, d\varepsilon \leq n\varepsilon_n^2.$$

[For the notation see, e.g., van der Vaart and Wellner (1996), Section 3.4.] This is partly a theoretical problem of getting a proof done, but is also due to an inherent instability of defining a procedure through a global search for a "peculiarity" of a contrast function (i.e., its point of extremum). The brackets are needed to have a better control of the contrast function, and the integral comes in, because the minimum contrast searches a continuous parameter set. The entropy criterion of Theorem 10.2 is simpler, because Le Cam's approach solves the problem of finding estimators in steps: it starts by selecting first a suitable net of approximations over the parameter set, and next picks the best element of this set by a testing argument. Recently, researchers have returned to this idea of using *sieves* [often credited to Grenander (1981), but present in Le Cam (1960a, 1969)], and it is likely to play an important role in methods for model selection.

The question arises if the general rates given by the preceding theorem are sharp. This is the case, under some conditions, as was shown by Birgé (1983). In contrast, the rates obtained by Le Cam for posterior distributions [Le Cam (1986), pages 509–529], using similar methods, appear to be sharp only under restrictive

dimensionality conditions. Recent work [Ghosal, Ghosh and van der Vaart (2000)] appears to indicate that this is because a completely metric entropy based approach is not feasible in this case, but an augmentation of Le Cam's methods does give the desired results.

11. Central dogmas of statistics. Le Cam was no fan of the method of maximum likelihood. His 1960 paper on local asymptotic normality offered an alternative class of estimators with the properties that maximum likelihood estimators were thought to have. He writes [Le Cam (1960a), page 94]:

> This author is firmly convinced that a recourse to maximum likelihood is justifiable only when one is dealing with families of distributions that are extremely regular. The cases in which m.l. estimates are easily obtainable and have been proved to have good properties are extremely restricted. One of the purposes of this paper is precisely to deëmphasize the role of m.l. estimates. Since, however, the m.l. estimates seem to exert a quasi-hypnotic attraction, a comparison of the results obtained herein with those obtainable for m.l. estimates is given below.

He kept this skepticism towards the centerpiece of Fisherian statistics throughout his career. His 1990 paper "Maximum likelihood: An introduction" [Le Cam (1990a)] is essentially a list of examples where maximum likelihood estimators do not exist, are not unique, or not consistent. His major book of 1986 hardly mentions the method of maximum likelihood. In a short section of his 1986 book the problems of the method are highlighted and Le Cam writes [Le Cam (1986), page 622]:

> The terms "likelihood" and "maximum likelihood" seem to have been introduced by R. A. Fisher who seems also to be responsible for a great deal of propaganda on the merits of the maximum likelihood method.
>
> In view of Fisher's vast influence, it is perhaps not surprising that the presumed superiority of the method is still for many an article of faith promoted with religious fervor. This state of affairs remains, in spite of a long accumulation of evidence to the effect that maximum likelihood estimates are often useless, or grossly misleading.

This is notable, as Le Cam's LAN theory is often viewed informally as showing that maximum likelihood estimators are asymptotically normal and efficient.

On the other hand, Le Cam contributed throughout his career to the investigation of conditions under which the method of maximum likelihood does work. This culminated in his paper Le Cam (1970b), which we discuss in Section 12.

Le Cam was milder towards Bayesian methods, but mostly because of the important role that Bayesian procedures play within the Wald decision theory. He writes [Le Cam and Yang (1990), page 165]:

> We have not taken a stand on Bayesianism as a philosophy. [...] Contrary to often expressed opinions, Bayes' approach had not disappeared from Statistics in the second quarter of the 20th Century. It was quite alive in most places except those that seem to have fallen under the influence of Fisher. It can certainly be used as shown here, but in practical situations it should be used with extreme caution.

The "extreme caution" should be understood in the context of the chapter, which contains a discussion of the inconsistency of many posteriors. Le Cam's relative preference of Bayesian over maximum likelihood methods does not reflect a truly Bayesian spirit. Decision theory, based on expectations over sample spaces, was central to his thinking, and priors featured in this as generators of procedures, not as expressions of subjective beliefs.

The word "likelihood" is not often used by Le Cam, and rarely in the way many statisticians write about "likelihood." In his discussion of the book by Berger and Wolpert (1988) on the likelihood principle, Le Cam discerns two bodies of theory. The type 1 statistical theory centers around an experiment, and the mathematical model for it, a set of probability measures. [The following quotes are from Le Cam's discussion, published in Berger and Wolpert (1988), pages 182–185.2.]

> This kind of endeavor has given us the Neyman–Pearson theory and Wald's theory of "statistical decision functions." One can readily claim that the whole enterprise is misguided, but it does seem to have a role to play in certain endeavors, like planning experiments, settling arguments that involve several scientists and odd questions such as "is methotrexate effective in the treatment of colon cancer."
>
> There is another body of theory, call it "type 2," that deals with axioms of coherent behavior and principles of evaluation of evidence. Some of it, and perhaps most of it, has to do with what "one" should "think" after the results of the experiments have become known.
>
> In a strictly mathematical view of the problem, there is no overlap between the two approaches because "type 1" does not have any probabilities to play with once the dice have been cast.
>
> Here the situation is complex because "type 1" theories have given proofs that "experiments" are characterized by the distributions of their likelihood functions. Also it is a standard result of "type 1" theories that Bayes procedures, or their limits form complete classes. A main difference is that the "type 1" theories insist that they are about risk functions, not possible interpretations of single posterior distributions.
>
> This author presumes that there is some value in some of "classical statistics" and also in the likelihood principle, but feels that one cannot support the practical application of either (or of other theories) on purely mathematical grounds. One should keep an open mind and be a bit "unprincipled."

Le Cam's most important objection to Berger and Wolpert's type 2 theory is that it is based on a notion of "evidence" attached to a pair (E, x) of an experiment and an observation, which is not clearly defined and cannot be used in arguments as if it were a mathematical function.

> I have, of course, no objection to a theory of "evidence" based on a function of pairs (E, x). It just fails to connect properly with my own intuitive notion of what evidence is. Therefore I do not feel bound in practice by the theorems derived from such a theory.
>
> In summary I remain opposed to the apparent normative aspect of a theory that says that I *must* abide by the LP when I am unable to put my emotion and various bits of knowledge, or lack of knowledge, into it.

12. Mathematical beauties. In this last section we draw attention to a few contributions, where Le Cam did not introduce a new statistical idea, but treated a classical subject with amazingly pretty mathematical techniques. Thus he showed that mathematical statistics does not have to consist of theorems whose statement, carefully including a long list of dreary regularity conditions, is longer than the proof.

First consider local asymptotic normality in the case of i.i.d. observations. As we noted the LAN expansion can be derived by a Taylor expansion under Cramér type conditions. This would involve at least two derivatives of the maps $\theta \mapsto p_\theta(x)$ and domination conditions on the second partial derivative. Some years after introducing LAN Le Cam found that the expansion is actually valid given a single derivative of the root density $\theta \mapsto p_\theta^{1/2}$.

Let the observations be an i.i.d. sample $X_1, \ldots, X_n$ from a density p_θ relative to some σ-finite measure μ on the measurable space $(\mathcal{X}, \mathcal{A})$ indexed by a parameter θ in an open subset $\Theta \subset \mathbb{R}^k$.

THEOREM 12.1. *Assume that there exists a measurable function $\dot{\ell}_\theta : \mathcal{X} \to \mathbb{R}^k$ such that, as $h \to 0$,*

$$\int \left[p_{\theta+h}^{1/2} - p_\theta^{1/2} - \tfrac{1}{2} h^T \dot{\ell}_\theta \, p_\theta^{1/2}\right]^2 d\mu = o(\|h\|^2). \tag{12.1}$$

Then the experiments $(P_\theta^n : \theta \in \Theta)$ are LAN (2.1) at θ with $\delta_n = n^{-1/2}$, $J_\theta = P_\theta \dot{\ell}_\theta \dot{\ell}_\theta^T$ and

$$\Delta_{n,\theta} := \frac{1}{\sqrt{n}} \sum_{i=1}^n \dot{\ell}_\theta(X_i).$$

Even though $\dot{\ell}_\theta$ is now defined through a derivative in quadratic mean, it must still be interpreted as the usual score function of the model, and J_θ as the usual Fisher information. This follows from the fact that, under regularity conditions, $\partial/\partial\theta \, p_\theta^{1/2} = \frac{1}{2}(\partial/\partial\theta \log p_\theta) \, p_\theta^{1/2}$. It is possible to give sufficient conditions for the differentiability of $\theta \mapsto p_\theta^{1/2}$ in terms of ordinary derivatives and domination or continuity conditions [see the appendix of Hájek (1972)], but the differentiability in quadratic mean is not only prettier, but also exactly right for getting the LAN expansion. Furthermore, it can be the basis for getting the asymptotic normality of maximum likelihood estimators or posterior distributions. (See below.)

The most remarkable feature of Theorem 12.1 is that it derives the quadratic, that is, second degree, expansion of the likelihood (2.1) from the existence of a single derivative of the map $\theta \mapsto p_\theta^{1/2}$. This is connected to the dual expression of the Fisher information as the variance of a score and minus the expectation of the derivative of the score, but is a good deal deeper. See Pollard's contribution to Le Cam's Festschrift [Pollard, Torgersen and Yang (1997)] for a further discussion.

The preceding theorem appears to be given for the first time in Le Cam (1966), embedded in a treatment of the more general situation of experiments consisting of observing a sample of independent variables (not necessarily asymptotically normal), and after preliminary work by Hájek in the special case of location models. The theorem is intimately connected to *Le Cam's second lemma*, obtained in Le Cam (1960a) and so named by Hájek and Šidák (1967). An extended version of this lemma specialized to i.i.d. observations [see Le Cam (1969), Proposition 1, page 47] asserts that, for a scaling rate δ_n such that $nH^2(\theta+\delta_n h_n, \theta) = O(1)$ for every converging sequence h_n, the approximation

$$\log \prod_{i=1}^{n} \frac{p_{\theta+\delta_n h}}{p_\theta}(X_i) = 2\sum_{i=1}^{n}\left(\sqrt{\frac{p_{\theta+\delta_n h}}{p_\theta}}(X_i) - 1\right) + \text{constant} + o_P(1)$$

can be valid only if the random variables on both left-hand and right-hand sides of this display are asymptotically normal. Such asymptotic normality (for every h) does not by itself imply LAN, because the random part of the limiting Gaussian process may not be linear in h, with the appropriate relationship to the quadratic part, as required by LAN. The linearization of the map $h \mapsto p^{1/2}_{\theta+\delta_n h}$ guaranteed by the differentiability (12.1) provides this linearity.

Rather than (12.1) one might require that, for some matrix J_θ, and uniformly in g, h running through compacta,

$$n\int \bigl(\sqrt{p_{\theta+\delta_n g}} - \sqrt{p_\theta}\bigr)\bigl(\sqrt{p_{\theta+\delta_n h}} - \sqrt{p_\theta}\bigr)\,d\mu \to \tfrac{1}{4} g J_\theta h. \tag{12.2}$$

Then it will follow that $\delta_n = \phi(n)/\sqrt{n}$ for a slowly varying function ϕ. If this function ϕ can be chosen equal to 1, (12.2) holds, and $nP_{\theta+\delta_n h}(p_\theta = 0) \to 0$ for all θ, then it can be deduced that (12.1) holds for almost every θ [see Le Cam (1969), pages 96–100, in particular the bottom of page 100]. One might conclude that for i.i.d. observations the differentiability (12.1) is intimately connected to the scaling rate $1/\sqrt{n}$.

The asymptotic normality of posterior distributions, known as the Bernstein–von Mises theorem, occupied Le Cam's attention as early as 1953, when he proved a version of this theorem. After explaining that the theorem is really due to Laplace, Le Cam and Yang [(1990), page 165] write:

> Fisher, whose work [(1922), (1925)] parallels that of Laplace in more than one way, does not seem to have added results on the behavior of posterior distributions. This may be because he did not view kindly the use of prior distributions and substituted a philosophy based on "fiducial probabilities." These seem to have been introduced as a result of a logically erroneous argument.
>
> Le Cam (1953) revived Laplace's argument. He used the convergence of posterior distributions to obtain a sort of asymptotic minimax theorem, and an asymptotic admissibility result for the one dimensional case. The conditions used there are very strong.

The "very strong conditions" were just the typical Taylor expansion type conditions in the spirit of Cramér. The following theorem is much nicer and can be deduced from Théorème 2 on pages 131–132 of Le Cam (1969), or from Chapter 7 in Le Cam and Yang (1990). We consider the i.i.d. case and make some simplifying conditions to bring out the essence of Le Cam's approach. It does apply more generally to locally asymptotically normal models, and can probably be used in other cases as well.

Let the observations be an i.i.d. sample $X_1, \ldots, X_n$ from a density p_θ relative to some σ-finite measure μ satisfying (12.1) at every point in a set $\Theta \subset \mathbb{R}^k$ and such that the maps $(\theta, x) \mapsto p_\theta(x)$ are measurable. We shall also assume that J_θ is nonsingular for every θ, that the map $\theta \mapsto J_\theta$ is continuous and that the parameter is *identifiable*, that is, the map $\theta \mapsto P_\theta$ is one-to-one.

The posterior density relative to a prior measure Π is given by

$$B \mapsto P_{\bar{\Theta}_n | X_1, \ldots, X_n}(B) = \frac{\int_B \prod_{i=1}^n p_\theta(X_i)\, d\Pi(\theta)}{\int \prod_{i=1}^n p_\theta(X_i)\, d\Pi(\theta)}.$$

THEOREM 12.2 (Laplace–Bernstein–von Mises–Le Cam). *Suppose that for some compact neighborhood $\Theta_0 \subset \Theta$ of θ_0, there exists a sequence of tests ϕ_n such that*

$$P_{\theta_0}^n \phi_n \to 0, \qquad \sup_{\theta \notin \Theta_0} P_\theta^n (1 - \phi_n) \to 0. \tag{12.3}$$

Furthermore, let the prior measure be absolutely continuous in a neighborhood of θ_0 with a continuous positive density at θ_0. Then the corresponding posterior distributions satisfy

$$\left\| P_{\sqrt{n}(\bar{\Theta}_n - \theta_0) | X_1, \ldots, X_n} - N(\Delta_{n,\theta_0}, J_{\theta_0}^{-1}) \right\| \overset{P_{\theta_0}^n}{\to} 0.$$

Again no second or higher order derivatives of p_θ relative to the parameter are needed. A further simplification of the theorem is obtained by assuming Θ to be compact. Then the condition (12.3) on existence of tests is trivially satisfied, because we can choose $\Theta_0 = \Theta$.

It is customary to write the Bernstein–von Mises theorem with a different "centering sequence." Under conditions somewhat stronger than we have imposed so far, the maximum likelihood estimators $\hat{\theta}_n$ satisfy

$$\sqrt{n}(\hat{\theta}_n - \theta) - \Delta_{n,\theta} \overset{P_\theta^n}{\to} 0.$$

If this is true, then by the invariance of the total variation norm under location and scale changes and the continuity of the normal distribution as a function of its mean value, the assertion of the theorem can be written

$$\left\| P_{\bar{\Theta}_n | X_1, \ldots, X_n} - N\left(\hat{\theta}_n, \frac{1}{n} J_{\theta_0}^{-1}\right) \right\| \overset{P_{\theta_0}^n}{\to} 0.$$

Whereas Wald would have written the maximum likelihood estimator, Le Cam would rather steer us away from this. The advantage of Le Cam's formulation is that maximum likelihood estimators need more regularity conditions for good behavior.

In Section 11 we noted that Le Cam was critical of the method of maximum likelihood, which in his view looks for a "peculiarity" of the likelihood function. This did not prevent him from investigating conditions under which the maximum likelihood estimator does have the properties that are usually ascribed to it. Le Cam's 1970 paper is a thorough investigation of the relationships between different sets of conditions related to the differentiability in quadratic mean of a density $\theta \mapsto p_\theta$. The discussion is complicated by the choice not to make the usual simplifying assumptions, such as an open parameter set, differentiability at every point of the parameter set, and positive and continuous Fisher information. Le Cam's (1970b) results on the asymptotic normality of the maximum likelihood estimator for parameters of dimension 2 or higher are interesting, but the most remarkable result concerns maximum likelihood estimators of one-dimensional parameters. His main result, Proposition 6 in the last section of Le Cam (1970b) has a long list of assumptions, but all of these are due to the choice to be as general as possible. The result is clearly a final one on the one-dimensional case.

The following theorem is a corollary of Le Cam's Proposition 6. Characteristically for Le Cam's writing, such a simple corollary is not included in his paper, with the result that Proposition 6 has been ignored by many authors in the past 30 years.

We consider again the setup of an i.i.d. sample from a density p_θ relative to some measure μ on some measurable space. The parameter set Θ is now assumed to be a sub-interval of the real line, which we shall first assume to be compact. Under the conditions of the following theorem there exist versions of the densities p_θ such that the processes $\theta \mapsto p_\theta(X_i)$ are separable. It is to be understood that the likelihood is constructed from such a version.

THEOREM 12.3. *Let $\Theta \subset \mathbb{R}$ be a compact interval and assume that the map $\theta \mapsto p_\theta^{1/2}$ is differentiable in quadratic mean* (12.1) *with continuous, positive Fisher information J_θ. Then if the data are sampled from p_θ for an interior point θ of Θ, there exist maximum likelihood estimators $\hat{\theta}_n$ and the sequence $\sqrt{n}(\hat{\theta}_n - \theta)$ is asymptotically normal with mean zero and variance $1/J_\theta$.*

The compactness assumption is unpleasant. Le Cam does not assume this, but assumes instead that the set $\{t : H^2(P_t, P_\theta) \le \varepsilon\}$ is bounded for some $\varepsilon > 0$ and, with the supremum taken over all partitions of Θ,

$$\sup_{\theta_0 < \theta_1 < \cdots < \theta_m} \sum_i H^2(P_{\theta_i}, P_{\theta_{i+1}}) < \infty.$$

This helps a little, but generally some sort of compactness assumption cannot be removed without addressing the problem of consistency first, possibly by different methods, not discussed by Le Cam (1970b).

Other relaxations allowed by Le Cam's Proposition 6 concern the differentiability in quadratic mean (for a result under θ it is not necessary to impose the differentiability throughout the interval), and the continuity and existence of the Fisher information [it suffices that $\lim_{t\to 0} H^2(P_{\theta+t}, P_\theta)/t^2$ exists in a neighborhood of the true θ and that this true θ is a Lebesgue point of the resulting function].

Acknowledgments. Rudy Beran, Lucien Birgé, David Pollard and Jon Wellner were very kind to read the present paper, to point out mistakes and to provide additional comments. I am very grateful for their support.

REFERENCES

BAHADUR, R. R. (1964). On Fisher's bound for asymptotic variances. *Ann. Math. Statist.* **35** 1545–1552.

BASAWA, I. V. and SCOTT D. J. (1983). *Asymptotic Optimal Inference for Nonergodic Models*. *Lecture Notes in Statist.* **17**. Springer, New York.

BEGUN, J. M., HALL, W. J., HUANG, W. M. and WELLNER, J. A. (1983). Information and asymptotic efficiency in parametric–nonparametric models. *Ann. Statist.* **11** 432–452.

BERGER, J. O. and WOLPERT, R. L. (1988). *The Likelihood Principle*, 2nd ed. IMS, Hayward, CA.

BICKEL, P. J., KLAASSEN, C. A. J., RITOV, Y. and WELLNER, J. A. (1993). *Efficient and Adaptive Estimation for Semiparametric Models*. Johns Hopkins Univ. Press.

BIRGÉ, L. (1983). Approximation dans les espaces métriques et théorie de l'estimation. *Z. Wahrsch. Verw. Gebiete* **65** 181–238.

BIRGÉ, L. (1986). On estimating a density using Hellinger distance and some other strange facts. *Probab. Theory Related Fields* **71** 271–291.

BIRGÉ, L. and MASSART, P. (1993). Rates of convergence for minimum contrast estimators. *Probab. Theory Related Fields* **97** 113–150.

BIRGÉ, L. and MASSART, P. (1998). Minimum contrast estimators on sieves: Exponential bounds and rates of convergence. *Bernoulli* **4** 329–375.

BLACKWELL, D. (1951). Comparison of experiments. *Proc. Second Berkeley Symp. Math. Statist. Probab.* 93–102. Univ. California Press, Berkeley.

BLACKWELL, D. (1953). Equivalent comparisons of experiments. *Ann. Math. Statist.* **24** 265–272.

BOHNENBLUST, H. F., SHAPLEY, L. S. and SHERMAN, S. (1949). Reconnaissance in game theory. *RAND Research Memorandum* **RM-208** 1–18.

BOLL, C. (1955). Comparison of experiments in the infinite case. Ph.D. dissertation, Dept. Statist., Stanford Univ.

BOURBAKI, N. (1955). *Espaces vectoriels topologiques*. Hermann, Paris.

BROWN, L. D. and LOW, M. G. (1996). Asymptotic equivalence of nonparametic regression and white noise. *Ann. Statist.* **24** 2384–2398.

CHERNOFF, H. (1956). Large sample theory: Parametric case. *Ann. Math. Statist.* **27** 1–22.

DONOHO, D. L. and LIU, R. C. (1991). Geometrizing rates of convergence. II, III. *Ann. Statist.* **19** 633–667, 668–701.

DUDLEY, R. M. (1967). The sizes of compact subsets of Hilbert spaces and continuity of Gaussian processes. *J. Funct. Anal.* **1** 290–330.

FISHER, R. A. (1922). On the mathematical foundations of theoretical statistics. *Philos. Trans. Roy. Soc. London Ser. A* **222** 309–368.

FISHER, R. A. (1925). Theory of statistical estimation. *Proc. Cambridge Philos. Soc.* **22** 700–725.
FISHER, R. A. (1935). *The Design of Experiments*. Oliver and Boyd, Edinburgh.
GHOSAL, S., GHOSH, J. K. and VAN DER VAART, A. W. (2000). Convergence rates of posterior distributions. *Ann. Statist.* **28** 500–531.
GRENANDER, U. (1981). *Abstract Inference.* Wiley, New York.
GUSHCHIN, A. A. (1995). Asymptotic optimality of parameter estimators under the LAQ condition. *Theory Probab. Appl.* **40** 261–272.
HÁJEK, J. (1962). Asymptotically most powerful rank-order tests. *Ann. Math. Statist.* **33** 1124–1147.
HÁJEK, J. (1970). A characterization of limiting distributions of regular estimates. *Z. Wahrsch. Verw. Gebiete* **14** 323–330.
HÁJEK, J. (1972). Local asymptotic minimax and admissibility in estimation. *Proc. Sixth Berkeley Symp. Math. Statist. Probab.* **1** 175–194. Univ. California Press, Berkeley.
HÁJEK, J. and ŠIDÁK, Z. (1967). *Theory of Rank Tests.* Academic Press, New York.
HEYER, H. (1973). *Mathematische Theorie statistischer Experimente*. Springer, Berlin.
IBRAGIMOV, I. A. and KHASMINSKII, R. Z. (1977). On the estimation of an infinite dimensional parameter in Gaussian white noise. *Soviet Math. Dokl.* **236** 1035–1055.
IBRAGIMOV, I. A. and HAS'MINSKII, R. Z. (1981). *Statistical Estimation*: *Asymptotic Theory*. Springer, New York.
JANSSEN, A., MILBRODT, H. and STRASSER, H. (1985). *Infinitely Divisible Statistical Experiments*. *Lecture Notes in Statist.* **27**. Springer, New York.
JEGANATHAN, P. (1982). On the asymptotic theory of estimation when the limit of the log likelihood ratios is mixed normal. *Sankhyā Ser. A* **44** 173–212.
JEGANATHAN, P. (1995). Some aspects of asymptotic theory with applications to time series models. *Econometric Theory* **11** 818–887.
KOLMOGOROV, A. N. and TIKHOMIROV, V. M. (1961). ε-entropy and ε-capacity of sets in function spaces. *Amer. Math. Soc. Trans. Ser. 2* **17** 277–364.
KOSHEVNIK, YU. A. and LEVIT, B. YA. (1976). On a nonparametric analogue of the information matrix. *Theory Probab. Appl.* **21** 738–753.
KOUL, H. L. and PFLUG, G. C. (1990). Weakly adaptive estimators in explosive autoregression. *Ann. Statist.* **18** 939–960.
LUXEMBURG, W. A. J. and ZAANEN, A. C. (1971). *Riesz Spaces.* North-Holland, Amsterdam.
MILLAR, P. W. (1979). Asymptotic minimax theorems for the sample distribution function. *Z. Wahrsch. Verw. Gebiete* **48** 233–252.
MILLAR, P. W. (1983). The minimax principle in asymptotic statistical theory. *École d'Été de Probabilités de St. Flour XI. Lecture Notes in Math.* **976** 76–267. Springer, New York.
MILLAR, P. W. (1985). Non-parametric applications of an infinite-dimensional convolution theorem. *Z. Wahrsch. Verw. Gebiete* **68** 545–556.
NEYMAN, J. (1934). On the two different aspects of the representative method: The method of stratified sampling and the method of purposive selection. *J. Roy. Statist. Soc.* **97** 558–625.
NUSSBAUM, M. (1996). Asymptotic equivalence of density estimation and Gaussian white noise. *Ann. Statist.* **24** 2399–2430.
PINSKER, M. S. (1980). Optimal filtering of square integrable signals in Gaussian white noise. *Problems Inform. Transmission* **16** 120–133.
PFANZAGL, J. and WEFELMEYER, W. (1982). *Contributions to a General Asymptotic Statistical Theory. Lecture Notes in Statist.* **13**. Springer, New York.
PFLUG, G. C. (1983). The limiting loglikelihood process for discontinuous density families. *Z. Wahrsch. Verw. Gebiete* **64** 15–35.
POLLARD, D., TORGERSEN, E. and YANG, G. L. (eds.) (1997). *Festschrift for Lucien Le Cam*: *Research Papers in Probability and Statistics.* Springer, New York.

PRAKASA RAO, B. L. S. (1968). Estimation of the location of the cusp of a continuous density. *Ann. Math. Statist.* **39** 76–87.

RAO, C. R. (1965). *Linear Statistical Inference and Its Applications*. Wiley, New York.

ROUSSAS, G. G. (1965). Asymptotic inference in Markov processes. *Ann. Math. Statist.* **36** 978–992.

ROUSSAS, G. G. (1972). *Contiguity of Probability Measures: Some Applications in Statistics.* Cambridge Univ. Press.

SCHAEFER, H. H. (1974). *Banach Lattices and Positive Operators.* Springer, New York.

SHERMAN, S. (1951). On a theorem of Hardy, Littlewood, Polya, and Blackwell. *Proc. Natl. Acad. Sci. U.S.A.* **37** 826–831.

STEIN, C. (1956). Inadmissibility of the usual estimator for the mean of a multivariate normal distribution. *Proc. Third Berkeley Symp. Math. Statist. Probab.* **1** 197–206. Univ. California Press, Berkeley.

STRASSER, H. (1985). *Mathematical Theory of Statistics*: *Statistical Experiments and Asymptotic Theory*. de Gruyter, Berlin.

TORGERSEN, E. (1968). Comparison of experiments when the parameter space is finite. Ph.D. dissertation, Dept. Statistics, Univ. California, Berkeley.

TORGERSEN, E. (1970). Comparison of experiments when the parameter space is finite. *Z. Wahrsch. Verw. Gebiete* **16** 219–249.

TORGERSEN, E. (1972). Comparison of translation experiments. *Ann. Math. Statist.* **43** 1383–1399.

TORGERSEN, E. (1991). *Comparison of Statistical Experiments.* Cambridge Univ. Press.

VAN DE GEER, S. (1993). Hellinger-consistency of certain nonparametric maximum likelihood estimators. *Ann. Statist.* **21** 14–44.

VAN DER VAART, A. W. (1994). Limits of experiments. Lecture notes, Yale Univ.

VAN DER VAART, A. W. and WELLNER, J. A. (1996). *Weak Convergence and Empirical Processes*. *With Applications to Statistics*. Springer, New York.

WALD, A. (1939). Contributions to the theory of statistical estimation and testing hypotheses. *Ann. Math. Statist.* **10** 299–326.

WALD, A. (1943). Tests of statistical hypotheses concerning several parameters when the number of observations is large. *Trans. Amer. Math. Soc.* **54** 426–482.

WALD, A. (1950). *Statistical Decision Functions.* Wiley, New York.

WONG, W. H. and SHEN, X. (1995). Probability inequalities for likelihood ratios and convergence rates of sieve MLE's. *Ann. Statist.* **23** 339–362.

DIVISION OF MATHEMATICS
AND COMPUTER SCIENCE
FREE UNIVERSITY
DE BOELELAAN 1081A
1081 HV AMSTERDAM
NETHERLANDS
E-MAIL: aad@cs.vu.nl